Briefings in Bioinformatics Advance Access published April 17, 2008

doi:10.1093/bib/bbn019

The relative value of operon predictions

Rutger W. W. Brouwer, Oscar P. Kuipers and Sacha A. F. T. van Hijum

Submitted: 11th January 2008; Received (in revised form): 21st March 2008

Abstract

For most organisms, computational operon predictions are the only source of genome-wide operon information. Operon prediction methods described in literature are based on (a combination of) the following five criteria: (i) intergenic distance, (ii) conserved gene clusters, (iii) functional relation, (iv) sequence elements and (v) experimental evidence. The performance estimates of operon predictions reported in literature cannot directly be compared due to differences in methods and data used in these studies. Here, we survey the current status of operon prediction methods. Based on a comparison of the performance of operon predictions on *Escherichia coli* and *Bacillus subtilis* we conclude that there is still room for improvement. We expect that existing and newly generated genomics and transcriptomics data will further improve accuracy of operon prediction methods.

Keywords: *operon; computational prediction; bioinformatics*

INTRODUCTION

Genes co-transcribed to polycistronic messenger RNAs are defined as operons, which are present in prokaryotes [1, 2]. Most operons are under the control of a single transcriptional promoter located upstream of the first gene of the operon. More complex transcriptional regulation with multiple promoters and transcriptional terminators in a single operon has also been reported [3].

It has been estimated that ~50% of genes in bacteria are located in operons [4], and several theories have been proposed to explain the formation of these transcriptional units [5, 6]. The first view is that operons evolved to ensure that genes are co-regulated [5]. This theory is supported by the observation that genes in operons often encode proteins that (i) are functionally related, such as enzymes catalyzing subsequent steps within metabolic pathways [7] or (ii) are members of a single protein complex [8].

The second view is the selfish operon model [6]. In this model, non-essential genes form operons via horizontal gene transfer to protect themselves from being removed from the genome. This view is based on the observation that numerous orthologous operons are conserved across bacterial and archaeal species [1, 9, 10].

Knowledge on the organization of genes in operons is used in many fields of prokaryotic research. Predicting the function of proteins is greatly aided by identifying operon structures, e.g. by applying the 'guilty by association' rule to remaining operon members when the function of one or more gene products is known [1, 9, 11]. Furthermore, operon information reduces the search space for determining *cis*-regulatory elements [12]. Operon information is also used to more reliably determine significant differential gene expression between experimental conditions in DNA microarray experiments [13, 14].

In the prokaryotic model organisms *Escherichia coli* and *Bacillus subtilis*, substantial numbers of operons have experimentally been verified [2, 15–17]. These collections of operons do not represent all the operons present in the genomes of these bacteria. To infer operon structures genome-wide in these and other prokaryotes, various computational methods have been developed (see below). Thus far, a comprehensive comparison of the results of these algorithms has not been performed. Here we compare, based on uniform criteria, the outcome of these prediction methods to experimentally verified operons for both *E. coli* and *B. subtilis*.

Corresponding authors. Oscar P. Kuipers and Sacha A. F. T. van Hijum, Department of Molecular Genetics, Groningen Biomolecular Sciences and Biotechnology Institute, Kerklaan 30, 9751 NN Haren, University of Groningen. E-mail: o.p.kuipers@rug.nl or s.a.f.t.van.hijum@rug.nl

Rutger W. W. Brouwer, Oscar P. Kuipers and Sacha A. F. T. van Hijum are all members of the Molecular Genetics Department of the University of Groningen. Currently, Sacha van Hijum is employed by the Interfacultary Centre of Functional Genomics, Ernst-Moritz-Arndt-Universität, Greifswald, Germany. The authors' focus is on the discovery and description of gene regulatory networks in Gram-positive bacteria by computational and experimental approaches.

COMPUTATIONAL OPERON PREDICTIONS

In recent years, various computational methods have been developed to infer operon structures in prokaryotes (Table 1) [7, 8, 10, 12, 18–45]. Implementations to predict operons for newly sequenced organisms are provided with only few of these studies [37, 42]. The results of most operon prediction methods are, however, made available by the original authors via the World Wide Web.

Five general features are used by operon prediction methods to predict operons: intergenic distance, conserved gene clusters, functional relation, sequence elements and experimental evidence (Table 1).

Intergenic spacing

The distance between open reading frames (ORFs) is a commonly used feature in the prediction of operons (Table 1). The intergenic distances between members of the same operon are relatively small as compared to those of genes not belonging to the same operon [4, 20]. Operons of which the members are highly expressed are the exceptions to this rule [4] since for these operons a wider gene spacing has been observed.

Conserved gene clusters

Conserved gene clusters have been widely used to predict operons with homologs present in the various sequenced genomes [4, 10]. Even among closely related species, gene order is rarely conserved [9, 16]. In the cases where this conservation does occur, the most common reason is that the genes are part of the same operon [10].

Functional relations

Genes in operons often have some kind of functional relation, such as their products being members of the same protein complex [8], or enzymes part of a single metabolic reaction pathway [7]. Operon prediction methods have therefore taken many functional classifications into account to exploit this property including Riley's functional annotation [46], metabolic pathways [7], clusters of orthologous groups of proteins (COG) [47] and gene ontologies (GO) [48]. All of these classifications can be used to determine functional relations between genes, and thus have prove valuable for the prediction of operons.

Genome sequence-based features

From the genome sequence of an organism several features have been obtained with which operons can be predicted. The presence of DNA motifs and other sequence elements such as transcriptional terminators [29, 49, 50], promoter sequences [18, 38] and transcription factor binding sites [45] have been used to predict operons. Recently a specific operon related DNA motif was proposed [42], the 'TTTTT' motif. This motif, of which the function is currently unknown, is overrepresented in the intergenic region of genes belonging to the same operon. Other indicators derived from the genome sequence include similarities in codon adaptation index between genes belonging to the same operon [36, 39].

Experimental evidence

Several studies have used gene-expression data derived from DNA microarray experiments to predict operons [23–25, 29, 31, 43]. Genes part of the same operon should show similar expression patterns. Therefore, correlations in gene expression in multiple DNA microarray experiments have been used to predict operon structure. However, perturbations in the expression of large numbers of genes in the DNA microarray experiments are required for such a methodology [23]. DNA microarray compendia querying a range of experimental conditions are therefore required to successfully apply this criterion to the prediction of operons.

Many methods have been explored to combine the prediction results of different features (Table 1). Salgado and coworkers used statistical log-likelihood scores [20–22, 38]. Other prediction methods have used Bayesian-based techniques [19, 29, 36, 40, 42], genetic algorithms [35] and machine-learning approaches [26, 39, 41, 44].

REPORTED PERFORMANCE OF OPERON PREDICTION METHODS

The performance of computational operon prediction methods is commonly estimated based on a comparison of their results to experimentally verified operons. Collections of verified operons are available for *E. coli* and *B. subtilis* [2, 15, 16]. However, the reported performances cannot be consistently compared because of several reasons.

Table 1: Properties of computational operon prediction methods

Authors	Year of publication	Feature					Scoring method
		Intergenic spacing	Conserved gene clusters	Functional relations	Genome sequence based	Experimental evidence	
Yada *et al.* [18]	1999	X			Promoters, transcriptional terminators, ribosome binding sites		Hidden Markov model
Craven *et al.* [19]	2000	X		Riley's functional classification	Promoters, transcriptional terminators, operon size	39 DNA microarray datasets	Naive Bayes
Salgado *et al.* [20]*	2000	X		Riley's functional classification			Log-likelihood scores
Ermolaeva *et al.* [10]*	2001		X				
Moreno-Hagelsieb and Collado-Vides [21]*	2002	X					Log-likelihood scores
Moreno-Hagelsieb and Collado-Vides [22]*	2002	X	X	Riley's functional classification			Log-likelihood scores
Sabatti *et al.* [23]	2002	X				72 DNA microarray datasets	Bayesian classifier
Tjaden *et al.* [24]	2002					Genome tilling DNA microarrays	
Zheng *et al.* [7]*	2002			Metabolic pathways			
Bockhorst *et al.* [25, 26]*	2003	X			Codon usage, promoters, transcriptional terminators, operon length	39 DNA microarray datasets	Bayesian network
Chen *et al.* [27, 28]	2004	X	X	COG	Transcriptional terminators, conserved promoters		Log-likelihood scores
de Hoon *et al.* [29]*	2004	X			Operon length	174 DNA microarray datasets	Bayesian classifier
Paredes *et al.* [30]	2004	X			Promoters, transcriptional terminators,		Emprical scoring scheme
Romero and Karp [8]	2004	X		Riley's functional classification, metabolic pathways, protein complex information, functional classification of upstream genes, similarity in codon usage			Log-likelihood scores
Steinhauser *et al.* [31]	2004	X				140 DNA microarray datasets	Unweighted average linkage-clustering algorithm
Wang *et al.* [32]	2004	X	X		Transcriptional terminators		Empirical scoring scheme
Yan *et al.* [12]	2004	X	X				
de Hoon *et al.* [33]	2005				Transcriptional terminators		

(continued)

Table 1: Continued

Authors	Year of publication	Feature					Scoring method
		Intergenic spacing	Conserved gene clusters	Functional relations	Genome sequence based	Experimental evidence	
Edwards *et al.* [34]*	2005	X	X				Maximum weighted maximum cardinality bipartite matching algorithm
Jacob *et al.* [35]*	2005	X	X	Metabolic pathways, protein function			Fuzzy guided genetic algorithm
Price *et al.* [36]*	2005	X	X	COG	Codon adaptation index		Naive Bayes approach
Westover *et al.* [37]	2005	X	X	Functional relatedness			Naïve Bayes approach
Janga *et al.* [38]*	2006				Oligo-nucleotide signatures		Log-likelihood scores
Zhang *et al.* [39]*	2006	X	X	Metabolic pathways, interacting protein domains			Support vector machine
Bergman *et al.* [40]*	2007	X	X				Bayesian hidden markov model
Charaniya *et al.* [41]	2007	X			Transcriptional terminators	67 DNA microarray datasets	Support vector machine
Dam *et al.* [42]*	2007	X	X	GO	TTTTT motif, gene length ratio		11 classifiers from PRTools Mathlab toolbox
Roback *et al.* [43]	2007	X				474 DNA microarray datasets	Logistic regression predictive model
Tran *et al.* [44]	2007	X		Metabolic pathways, GO			Neural network incorporating the criteria combined with results from [28, 36, 37]
Laing *et al.* [45]	2008				Transcription factor binding sites		

A list of all the operon predictions methods described in literature together with the one which they base their predictions. These features can be roughly divided into five categories: intergenic spacing, conserved gene clusters, functional relations, genome sequence based and experimental evidence. The last column describes which method was used to decide based on the feature scores which genes form operons. Operon prediction methods of which the performances were estimated are marked with '*'.

Firstly, the verified operons used to estimate performances may differ between studies. For *B. subtilis*, three different collections of verified operons are available, namely, the Itoh collection [16], operon database (ODB) [2] and the DBTBS database [17]. The DBTBS database contains the most recent collection of experimentally verified operons for *B. subtilis*. It has thus far not been used in the validation of operon prediction methods but does list the experimental evidence used to identify operons. For *E. coli* verified operons are commonly obtained from the RegulonDB database [15]. This database is updated regularly.

Secondly, several different methods have been used to estimate the performance of operon predictions. Most methods to estimate the performances of operon predictions are based on gene pairs. Salgado and coworkers [20] used the fraction of within operon (WO) gene pairs correctly predicted (true positives, TP) as a measure of sensitivity. As a measure of specificity they determined the fraction of correctly predicted gene pairs at the operon boundaries (true negatives, TN; transcriptional unit boundary pair, TUB) [20]. Another method used by Craven and coworkers [19] uses the same sensitivity measure as the estimates from Salgado and coworkers. However, specificity is based on the number of WO gene pairs not predicted (false positives, FP) [19]. Variations on these methods to estimate performance have been used in most literature proposing operon prediction methods.

Finally, operon prediction methods have been developed to predict a specific subset of operons for a given genome. An example is the method developed by Zheng and coworkers [7]. This method is meant to predict operons which the members encode enzymes catalyzing subsequent steps in metabolic pathways. The performance estimate reported by the authors is thus based on a limited number of operon structures.

We have estimated the genome-wide performances of several operon predictions for the model organisms *E. coli* and *B. subtilis* (see below) based on uniform criteria and a single set of experimentally verified operons. Only operon predictions with results available online were used. In the cases where thresholds needed to be applied the parameters and/or thresholds reported to yield optimal operon predictions by the respective authors were used.

COMPARING THE PERFORMANCE OF OPERON PREDICTIONS

To compare operon predictions, their concordance to verified operons of *E. coli* and *B. subtilis* was determined using the sensitivity and specificity measure which is based on WO and TUB gene pairs (see above, Figure 1). However, this measure might not reflect how well operon prediction methods predict complete operons. Therefore the percentage of correctly predicted verified operons has also been determined for the respective operon predictions (Figure 2).

The goal of the analysis presented here (Figures 1 and 2) is to determine the performances of operon predictions based on all the verified operons in *E. coli* and *B. subtilis*. Alternative transcripts in operons and single-gene transcriptional units were not incorporated in our performance analyses, since most operon predictions do not list either of these. The collections of experimentally verified operons were obtained from RegulonDB (*E. coli*) [15] and DBTBS (*B. subtilis*) [17]. From DBTBS only operon structures verified by northern analyses were used.

In both the gene pair and the operon-based analyses performed in this study, the best performance is obtained by the prediction performed by Dam and coworkers [42] (Figures 1 and 2). Their prediction method takes into account multiple criteria (Table 1) among which the presence of a 'TTTTT' DNA motif in the intergenic space between genes. The reported sensitivity and specificity for *E. coli* of the prediction described by Dam and coworkers [42] was 90 and 94%, respectively. These are higher than our estimates of 87 and 82% (Figure 1). The authors do report however, that the performance of their method decreases by 12% for organisms other than *E. coli* and *B. subtilis* [42].

Several operon predictions exhibit low specificity and sensitivity scores in our analysis, such as the prediction of Zheng and coworkers [7] and Ermolaeva and coworkers [10]. These operon predictions have been reported to only accurately predict a subset of the operons present in the genome. The prediction method developed by Ermolaeva and coworkers, for example, specifically predicts operons preserved in the 39 genomes used in their analysis. Those operons of which the structure is not preserved across these organisms are not expected to be predicted by this method [10].

Both the operon prediction performed by Salgado and coworkers [20] as well as the operon predictions performed by Moreno-Hagelsieb and

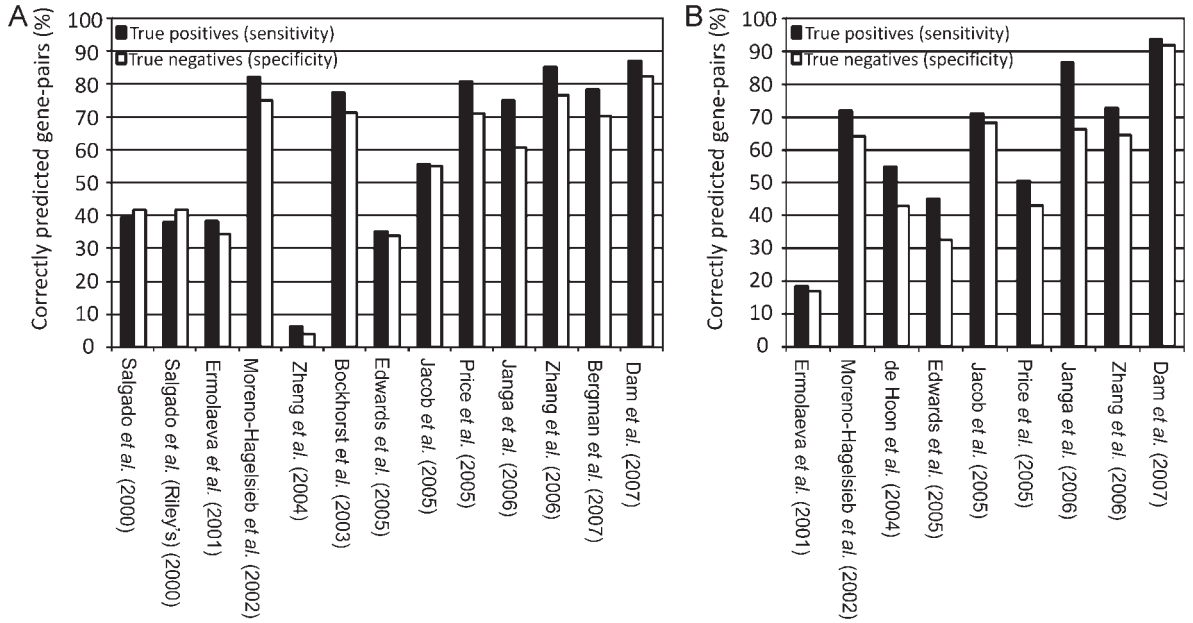

Figure 1: The estimated sensitivity and specificity of operon predictions for *E. coli* and *B. subtilis*. The sensitivities and specificities of operon predictions based on verified operons from RegulonDB [15] for *E. coli* (**A**) and DBTBS [17] for *B. subtilis* (**B**). True positive percentage is defined as the percentage of gene-pairs correctly predicted to be in operons divided by the total number of gene-pairs in operons and serves as a measure of sensitivity. True negative percentage is the percentage of gene pairs correctly predicted at the boundaries of operons divided by their total number and is a measure of specificity of operon predictions. The operon predictions and details of the analysis are available as supplementary data.

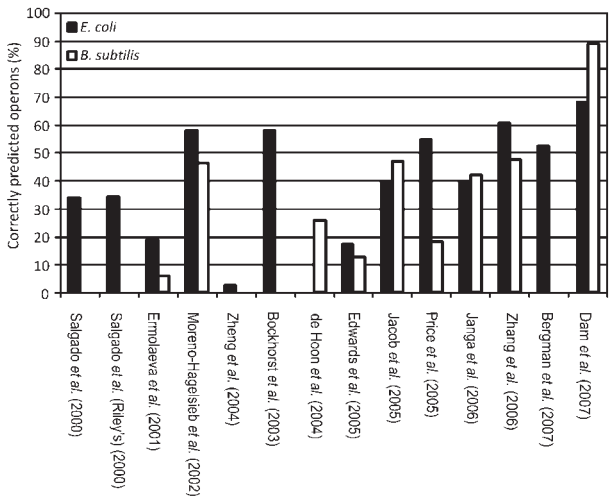

Figure 2: The performance of operon predictions using complete operons. The performances of operon predictions determined based on complete operons for *E. coli* and *B. subtilis*. The performance is defined as the percentage of verified operons correctly predicted by each of the operon prediction methods. The experimentally verified operons were obtained from RegulonDB for *E. coli* [15] and DBTBS [17] for *B. subtilis*. The operon predictions and details of the analysis are available as supplementary data.

Collado-Vides [21] use the same method to predict operons. However, a large difference in the performances between these operon predictions was observed (for whole operons 24%, Figure 2). We hypothesize that the larger number of verified operons available to the more recent prediction by Moreno-Hagelsieb and Collado-Vides [21] allowed increasing the WO pair performance of their method from 38% to 82% (Figure 1).

In contrast, the performances of the same operon prediction methods applied to *E. coli* and *B. subtilis* are similar (Figures 1 and 2). For the predictions performed by Jacob and coworkers [35] and Price and coworkers [36], this general observation does not hold true. The performance of the method performed by Jacob and coworkers is much better for *B. subtilis* than for *E. coli*, as opposed to that of Price and coworkers (Figure 2). The prediction performed by Price and coworkers was based on verified operons assembled by Itoh and coworkers [16] for *B. subtilis*. We based our analysis on the operons verified by northern blot analyses from DBTBS which may account for the differences in performance.

Generally, operon prediction methods show substantially lower scores when dealing with entire

operons as opposed to gene pairs (Figures 1 and 2). These lower scores are to be expected, since an operon has two TUBs and at least one WO. Therefore, one can calculate the entire operon score as a weighted product of the sensitivity and the specificity scores. Both methods to estimate the performance of operon predictions show similar results (Figures 1 and 2). In both analyses the best scoring prediction was that developed by Dam and coworkers [42].

CONCLUSIONS

The performance estimates of computational operon prediction methods reported in literature cannot reliably and systematically be compared. Therefore we re-estimated these performances in a single analysis based on WO and TUB gene pairs as measures of sensitivity and specificity. We observed that one of the eldest operon predictions performed by Moreno-Hagelsieb and Collado-Vides [21] using only intergenic distance outperforms many of the more recent predictions for both *E. coli* and *B. subtilis*. This observation emphasizes the power of using intergenic distance in the prediction of operons. The best performing prediction was performed by Dam and coworkers [42].

Dam and coworkers [42] reported that larger collections of verified operons do not significantly improve the results of their prediction. Other sources of genomics data may, however, still improve their accuracy. For example, new genome sequences are becoming available regularly. More sequence information may greatly improve the predictive value of conserved gene clusters [10]. Another improvement is possible in the use of DNA microarray data. Sabatti and coworkers [23] performed their operon prediction based on 72 DNA microarray datasets for *E. coli* (Table 1). At present data from many more DNA microarray experiments is available for various organisms in online databases such as Gene Expression Omnibus [51], ArrayExpress [52] and Stanford DNA microarray database [53], which will surely give rise to still better operon definitions when combined with appropriate computational prediction methods.

Key Points

- The criteria on which all operon prediction methods base their predictions can be divided in five classes.
- The performance estimates of operon predictions as stated in literature cannot be directly compared.
- We compared the concordance of a number of operon predictions to experimentally verified operons of *E. coli* and *B. subtilis*.

Acknowledgements

This work is part of the BioRange programme of the Netherlands Bioinformatics Centre (NBIC), which is supported by a BSIK grant through the Netherlands Genomics Initiative (NGI). Part of the work was supported by EU FW6 grant Bacell health contract number LSHC-CT-2004-503468. Netherlands Bioinformatics Centre (BioRange SP3.7.4 to R.W.W.B.). European union framework 6 (LSHC-CT-2004-503468 to O.P.K.).

SUPPLEMENTARY DATA

Additional files and figures are contained in the supplementary web site: http://bioinformatics.biol.rug.nl/supplementary/operon_data.

References

1. Wolf YI, Rogozin IB, Kondrashov AS, *et al*. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res* 2001;**11**:356–72.
2. Okuda S, Katayama T, Kawashima S, *et al*. ODB: a database of operons accumulating known operons across multiple genomes. *Nucleic Acids Res* 2006;**34**:D358–62.
3. Okuda S, Kawashima S, Kobayashi K, *et al*. Characterization of relationships between transcriptional units and operon structures in *Bacillus subtilis* and *Escherichia coli*. *BMC Genomics* 2007;**8**:48.
4. Price MN, Arkin AP, Alm EJ. The life-cycle of operons. *PLoS Genet* 2006;**2**:e96.
5. Price MN, Huang KH, Arkin AP, *et al*. Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Res* 2005;**15**:809–19.
6. Lawrence JG, Roth JR. Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* 1996;**143**: 1843–60.
7. Zheng Y, Szustakowski JD, Fortnow L, *et al*. Computational identification of operons in microbial genomes. *Genome Res* 2002;**12**:1221–30.
8. Romero PR, Karp PD. Using functional and organizational information to improve genome-wide computational prediction of transcription units on pathway-genome databases. *Bioinformatics* 2004;**20**:709–17.
9. Siefert JL, Martin KA, Abdi F, *et al*. Conserved gene clusters in bacterial genomes provide further support for the primacy of RNA. *J Mol Evol* 1997;**45**:467–72.
10. Ermolaeva MD, White O, Salzberg SL. Prediction of operons in microbial genomes. *Nucleic Acids Res* 2001;**29**: 1216–21.
11. Overbeek R, Fonstein M, D'Souza M, *et al*. The use of gene clusters to infer functional coupling. *Proc Natl Acad Sci USA* 1999;**96**:2896–901.
12. Yan B, Methe BA, Lovley DR, *et al*. Computational prediction of conserved operons and phylogenetic footprinting of transcription regulatory elements in the metal-reducing bacterial family *Geobacteraceae*. *J Theor Biol* 2004; **230**:133–44.

13. Price MN, Arkin AP, Alm EJ. OpWise: operons aid the identification of differentially expressed genes in bacterial microarray experiments. *BMC Bioinformatics* 2006;**7**:19.
14. Carpentier AS, Riva A, Tisseur P, *et al*. The operons, a criterion to compare the reliability of transcriptome analysis tools: ICA is more reliable than ANOVA, PLS and PCA. *Comput Biol Chem* 2004;**28**:3–10.
15. Salgado H, Gama-Castro S, Peralta-Gil M, *et al*. RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res* 2006;**34**:D394–97.
16. Itoh T, Takemoto K, Mori H, *et al*. Evolutionary instability of operon structures disclosed by sequence comparisons of complete microbial genomes. *Mol Biol Evol* 1999;**16**:332–46.
17. Sierro N, Makita Y, de Hoon M, *et al*. DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. *Nucleic Acids Res* 2008;**36**:D93–6.
18. Yada T, Nakao M, Totoki Y, *et al*. Modeling and predicting transcriptional units of Escherichia coli genes using hidden Markov models. *Bioinformatics* 1999;**15**:987–93.
19. Craven M, Page D, Shavlik J, *et al*. A probabilistic learning approach to whole-genome operon prediction. *Proc Int Conf Intell Syst Mol Biol* 2000;**8**:116–27.
20. Salgado H, Moreno-Hagelsieb G, Smith TF, *et al*. Operons in *Escherichia coli*: Genomic analyses and predictions. *Proc Natl Acad Sci USA* 2000;**97**:6652–7.
21. Moreno-Hagelsieb G, Collado-Vides J. A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics* 2002;**18**:S329–36.
22. Moreno-Hagelsieb G, Collado-Vides J. Operon conservation from the point of view of *Escherichia coli*, and inference of functional interdependence of gene products from genome context. *In Silico Biol* 2002;**2**:87–95.
23. Sabatti C, Rohlin L, Oh MK, *et al*. Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Res* 2002;**30**:2886–93.
24. Tjaden B, Saxena RM, Stolyar S, *et al*. Transcriptome analysis of *Escherichia coli* using high-density oligonucleotide probe arrays. *Nucleic Acids Res* 2002;**30**:3732–8.
25. Bockhorst J, Qiu Y, Glasner J, *et al*. Predicting bacterial transcription units using sequence and expression data. *Bioinformatics* 2003;**19**(Suppl. 1):i34–43.
26. Bockhorst J, Craven M, Page D, *et al*. A Bayesian network approach to operon prediction. *Bioinformatics* 2003;**19**:1227–35.
27. Chen X, Su Z, Xu Y, *et al*. Computational prediction of operons in *Synechococcus* sp. WH8102. *Genome Inform* 2004; **15**:211–22.
28. Chen X, Su Z, Dam P, *et al*. Operon prediction by comparative genomics: an application to the *Synechococcus* sp. WH8102 genome. *Nucleic Acids Res* 2004;**32**:2147–57.
29. De Hoon MJ, Imoto S, Kobayashi K, *et al*. Predicting the operon structure of *Bacillus subtilis* using operon length, intergene distance, and gene expression information. *Pac Symp Biocomput* 2004;276–87.
30. Paredes CJ, Rigoutsos I, Papoutsakis ET. Transcriptional organization of the *Clostridium acetobutylicum* genome. *Nucleic Acids Res* 2004;**32**:1973–81.
31. Steinhauser D, Junker BH, Luedemann A, *et al*. Hypothesis-driven approach to predict transcriptional units from gene expression data. *Bioinformatics* 2004;**20**:1928–39.
32. Wang L, Trawick JD, Yamamoto R, *et al*. Genome-wide operon prediction in *Staphylococcus aureus*. *Nucleic Acids Res* 2004;**32**:3689–702.
33. De Hoon MJ, Makita Y, Nakai K, *et al*. Prediction of transcriptional terminators in *Bacillus subtilis* and related species. *PLoS Comput Biol* 2005;**1**:e25.
34. Edwards MT, Rison SCG, Stoker NG, *et al*. A universally applicable method of operon map prediction on minimally annotated genomes using conserved genomic context. *Nucleic Acids Res* 2005;**33**:3253–62.
35. Jacob E, Sasikumar R, Nair KN. A fuzzy guided genetic algorithm for operon prediction. *Bioinformatics* 2005;**21**: 1403–7.
36. Price MN, Huang KH, Alm EJ, *et al*. A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res* 2005;**33**:880–92.
37. Westover BP, Buhler JD, Sonnenburg JL, *et al*. Operon prediction without a training set. *Bioinformatics* 2005;**21**: 880–8.
38. Janga SC, Lamboy WF, Huerta AM, *et al*. The distinctive signatures of promoter regions and operon junctions across prokaryotes. *Nucleic Acids Res* 2006;**34**: 3980–7.
39. Zhang GQ, Cao ZW, Luo QM, *et al*. Operon prediction based on SVM. *Comput Biol Chem* 2006;**30**:233–40.
40. Bergman NH, Passalacqua KD, Hanna PC, *et al*. Operon Prediction for sequenced bacterial genomes without experimental information. *Appl Environ Microbiol* 2007;**73**: 846–54.
41. Charaniya S, Mehra S, Lian W, *et al*. Transcriptome dynamics-based operon prediction and verification in *Streptomyces coelicolor*. *Nucleic Acids Res* 2007;**35**:7222–36.
42. Dam P, Olman V, Harris K, *et al*. Operon prediction using both genome-specific and general genomic information. *Nucleic Acids Res* 2007;**35**:288–98.
43. Roback P, Beard J, Baumann D, *et al*. A predicted operon map for *Mycobacterium tuberculosis*. *Nucleic Acids Res* 2007;**35**: 5085–95.
44. Tran TT, Dam P, Su Z, *et al*. Operon prediction in *Pyrococcus furiosus*. *Nucleic Acids Res* 2007;**35**:11–20.
45. Laing E, Sidhu K, Hubbard SJ. Predicted transcription factor binding sites as predictors of operons in *Escherichia coli* and *Streptomyces coelicolor*. *BMC Genomics* 2008;**9**:79.
46. Riley M. Functions of the gene products of *Escherichia coli*. *Microbiol Mol Biol Rev* 1993;**57**:862–952.
47. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. *Science* 1997;**278**:631–7.
48. Ashburner M, Ball CA, Blake JA, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;**25**:25–9.
49. Ermolaeva MD, Khalak HG, White O, *et al*. Prediction of transcription terminators in bacterial genomes. *J Mol Biol* 2000;**301**:27–33.
50. Kingsford CL, Ayanbule K, Salzberg SL. Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol* 2007;**8**:R22.
51. Barrett T, Troup DB, Wilhite SE, *et al*. NCBI GEO: mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Res* 2007;**35**:D760–5.

52. Parkinson H, Kapushesky M, Shojatalab M, *et al*. Array Express—a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res* 2007;**35**: D747–50.

53. Demeter J, Beauheim C, Gollub J, *et al*. The Stanford Microarray Database: implementation of new analysis tools and open source release of software. *Nucleic Acids Res* 2007; **35**:D766–70.