

# PFP: A Computational Framework for Phylogenetic Footprinting in Prokaryotic Genomes

Dongsheng Che<sup>1,2</sup>, Guojun Li<sup>1</sup>, Shane T. Jensen<sup>3</sup>, Jun S. Liu<sup>4</sup>, and Ying Xu<sup>1</sup>

<sup>1</sup> Computational Systems Biology Laboratory,  
Department of Biochemistry and Molecular Biology and Institute of Bioinformatics,  
University of Georgia, Athens, GA 30602, USA

<sup>2</sup> Department of Computer Science, University of Georgia, Athens, GA 30602, USA

<sup>3</sup> Department of Statistics, The Wharton School, University of Pennsylvania,  
Philadelphia, PA 19104, USA

<sup>4</sup> Department of Statistics, Harvard University, Cambridge, MA 02138, USA

**Abstract.** Phylogenetic footprinting is a widely used approach for the prediction of transcription factor binding sites (TFBSs) through identification of conserved motifs in the upstream sequences of orthologous genes in eukaryotic genomes. However, this popular strategy may not be directly applicable to prokaryotic genomes, where typically about half of the genes in a genome form multiple-gene transcription units or operons. The promoter sequences for these operons are located in the inter-operonic rather than inter-genic regions, which require prediction of TFBSs at the transcriptional unit instead of individual gene level. We have formulated as a bipartite graph matching problem the identification of conserved operons (including both single-gene and multi-gene operons) whose individual gene members are orthologous between two genomes and present a graph-theoretic solution. By applying this method to *Escherichia coli* K12 and 11 of its phylogenetically neighboring species, we have predicted 2,478 sets of conserved operons, and discovered potential binding motifs for each of these operons. By comparing the prediction results of our approach and other prediction approaches, we conclude that it is advantageous to use our approach for prediction of *cis* regulatory binding sites in prokaryotes. The prediction software package PFP is available at <http://csbl.bmb.uga.edu/~dongsheng/PFP>.

## 1 Introduction

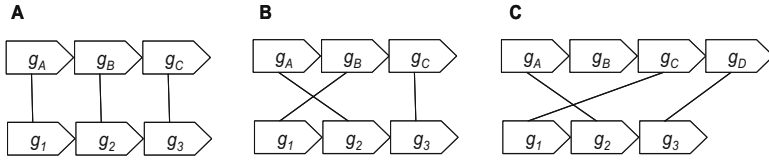
Phylogenetic footprinting is a method for identification of *cis* regulatory elements in promoter regions of orthologous genes across species [1]. This strategy attempts to find conserved sequence motifs in the provided promoter regions based on the assumption that functional elements, such as transcription factor binding sites, evolve more slowly than non-functional elements over time. A prerequisite for using a phylogenetic footprinting approach is the mapping of orthologous genes across multiple genomes (often called *reference* genomes).

A number of orthology mapping approaches, mainly sequence similarity-based such as COG [2] and OrthoMCL [3], have been widely used. By applying such orthology mapping methods to eukaryotic genomes, a number of research groups have carried out studies on identification of *cis* regulatory motifs at a genome scale. For example, Wang *et al.* [4] developed PhyloNet to search for regulatory motifs in *Saccharomyces cerevisiae* by using three other yeast genomes as reference genomes and identified more than 90% of the known TFBSs in *Saccharomyces cerevisiae*. Using several mammalian genomes as references, Xie *et al.* [5] successfully identified a number of transcription regulatory motifs in the human genome.

A similar phylogenetic footprinting strategy may not be directly applicable to prokaryotic genomes due to their different genomic structures from the eukaryotic ones. Typically about half of the genes in a prokaryotic genome are *polycistronic*, *i.e.*, they are organized into multi-gene transcriptional units (or multi-gene operons), genes of each of which share a common promoter and terminator. Multi-gene operons add a new challenge to the identification problem of orthologous promoter regions: promoters are associated with operons rather than individual genes and may not necessarily be conserved across multiple genomes. Thus, relationships between operons across genomes are more complex in general than those between orthologous genes. In addition, the sequence similarity-based approach cannot correctly characterize orthologous relationships in some cases. For prokaryotes, the true orthology can be elucidated by deriving conserved operons across multiple genomes. This is because that homologous genes are more likely to be orthologous if their neighboring genes within an operon are also homologous [6].

Numerous computational methods have been developed to predict operons in prokaryotic genomes, including OFS [7], OPERON [8], OperonDT [9], VIMSS [10], and UNIPOP (manuscript submitted). The prediction accuracy of the best programs has reached 90% on several model genomes such as *E. coli* and *Bacillus subtilis* [11]. It has been previously observed that “conserved” operons may only have their gene list conserved but not necessarily the gene order within the list. In this study, we consider both cases: *category-1* for conserved operons with both conserved gene list and order and *category-2* for conserved operons with only conserved gene list (Figure 1A and 1B). In addition, we also have considered *category-3* for partially conserved operons, which is defined as follows: two operons from different genomes are partially conserved if they have at least one pair of orthologous genes (Figure 1C). Clearly, the multiple scenarios of operon conservation complicate the derivation of orthologous upstream sequences for the purpose of phylogenetic footprinting analysis in prokaryotic species.

Previous work on extracting promoter sequences of orthologous genes for phylogenetic footprinting analysis has been done in a simplistic manner. Basically, orthologous genes are collected using sequence similarity-based approaches, then the intergenic sequences of individual genes with the upstream region of its predicted operon are concatenated [12,13,14]. This strategy has also been used in a recent computational tool ‘microFootPrinter’ [15]. To address the issue of



**Fig. 1.** Three categories of operon conservation. Boxes represent genes and consist of an operon. Lines indicate sequence similarity between two genes. (A) Conserved with both gene list and order; (B) Conserved with gene list only; and (C) Partially conserved.

including upstream sequences for internal genes in an operon, Jensen *et al* [16] considered only the “promoter” regions of genes with upstream intergenic regions longer than 50 bp (called *beginning* genes of an operon). This approach is also problematic since it considers only operons that have both conserved gene list and gene order. There remains a need for more careful and more accurate treatment of the “corresponding” promoters of orthologous genes in prokaryotes.

In this paper, we derive conserved operons among multiple genomes for phylogenetic footprinting analysis and provide a superior treatment of promoter regions of orthologous genes. To fully consider all operons with different levels of evolutionary conservation, we designed an algorithm, *OPERMAP*, to find operons across reference genomes. By applying this algorithm, we have identified 2,478 *E. coli* operons that are conserved across multiple (reference) genomes. In addition, we have developed a pipeline consisting of multiple motif discovery programs for the prediction of conserved sequence motifs. Performance comparison on known binding sites of *E. coli* suggests that our approach tend to generate more reliable orthologous promoter regions (*i.e.*, regions containing the binding sites for orthologous TFs) than previous approaches for motif finding at the genome scale in prokaryotes.

## 2 Methods

We divide our procedure of phylogenetic footprinting in prokaryotes into five steps:

1. Selecting reference genomes for a target genome;
2. Predicting operons of all selected genomes at genome-scale;
3. Predicting conserved operons across selected genomes;
4. Obtaining promoter sequences of conserved operons;
5. Predicting binding sites using our motif-finding pipeline.

Below, we present the details of each step.

**Reference Genome Selection.** Selecting suitable reference genomes for comparison to the target genome of interest is a key step in the phylogenetic footprinting process. A candidate reference genome should be phylogenetically close to the target genome. A large list of candidate genomes is not essential since using

a large number of genomes for motif discovery does not seem to improve performance [17]. This has also been observed in our experiments (data not shown). Accordingly, our selection strategy is to choose 10-15 reference genomes belonging to the same class with similar genome sizes to that of the target genome.

In this study, *E. coli* K12 is our target genome and 11 other  $\gamma$ -proteobacteria were chosen as reference genomes. The names and genome sizes of 12 genomes are listed as follows: *Aeromonas hydrophila* ATCC\_7966 (4.6 Mb), *Erwinia carotovora atroseptica* SCRI1043 (4.9 Mb), *E. coli* K12 (4.5 Mb), *Photobacterium profundum* SS9 (6.3 Mb), *Photorhabdus luminescens* (5.6 Mb), *Pseudomonas fluorescens* Pf-5 (6.9 Mb), *Salmonella enterica* Choleraesuis (4.9 Mb), *Shewanella ANA 3* (5.2 Mb), *Shigella sonnei* Ss046 (4.9 Mb), *Sodalis glossinidius morsitans* (4.2 Mb), *Vibrio parahaemolyticus* (5.1 Mb) and *Yersinia pestis* Antiqua (4.8 Mb).

**Operon Prediction.** For each of the selected genomes, operon prediction at the genome scale is performed using our own program UNIPOPOP (manuscript submitted). We choose UNIPOPOP because it outperforms other operon programs in terms of prediction accuracy. In addition, unlike most of operon programs, UNIPOPOP does not need extra feature information (*i.e.*, gene function annotation), which is not available for newly sequenced genomes. The key idea of UNIPOPOP is to predict operons through identification of conserved gene clusters across multiple genomes. Briefly, given a target genome and  $N$  reference genomes, we predict  $N$  versions of operon maps for the target genome by comparing and deriving conserved gene clusters between the target genome and each of the reference genomes. We consider two sets of contiguous genes from two genomes to be conserved gene clusters (or operons) if the following conditions are satisfied: a). Each member of a gene cluster is transcribed in the same direction; b). The total intergenic distance within each group is less than the maximum allowed distance (*MAD*); c). The number of mappings of homologous gene pairs between two groups is at least two. We then obtain a consensus version of operon map using a voting scheme on  $N$  versions of operon maps. In this study, operon structures for each of the 12 genomes were predicted by using 348 reference genomes from the NCBI GenBank database (<http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi>).

**Identification of Conserved Operons.** Having predicted operon structures for the 12 species, we need to identify “orthologous” operons among these prokaryotes. We have developed an algorithm, called *OPERMAP*, to identify the corresponding conserved operon in a particular reference genome for a given query operon in the target genome. We now describe the *OPERMAP* approach in detail as follows.

The input to the algorithm consists of

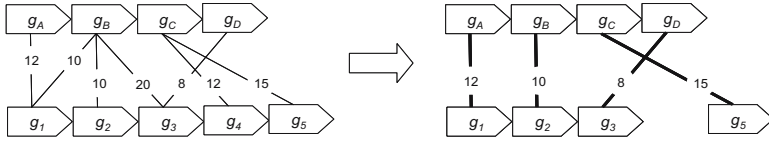
1. a query operon  $U$  in the target genome,
2. a collection of all predicted operons  $[V_1, V_2, \dots, V_k]$  in the reference genome, and
3. a threshold for the degree of conservation (*TDC*) between two operons.

The output of the program is the operon pair  $(U, V^*)$  between the query operon  $U$  and the best conserved operon  $V^*$  from the reference genome. The algorithm proceeds as follows:

1. Calculate the degree of conservation between query operon  $U$  and each candidate operon  $[V_1, V_2, \dots, V_k]$  in the reference genome.
  - (a) For each operon  $V_i \in [V_1, V_2, \dots, V_k]$ , construct a bipartite graph  $G_i = (U, V_i, E_i)$ , where all the genes in  $U$  and all the genes in the  $i$ -th operon  $V_i$  are represented as vertices. A pair of genes is considered to be *homologous* if their reciprocal BLAST e-values are both  $< 10^{-6}$ , and a homologous relationship between a gene in  $U$  and a gene in  $V_i$  is represented by an edge in  $E_i$ . The weight of each edge in  $E_i$  is set to be the average of  $-\log(\text{e-value})$  of the BLAST between the pair of genes.
  - (b) Calculate the maximum weighted maximum cardinality bipartite matching (*mwmcm*)  $M_i$  on each graph  $G_i$ , in a similar fashion to that of [18]. Each matched edge in *mwmcm* reflects the orthology relationship between the pair of genes.
  - (c) Calculate the degree of conservation  $DC_i = |M_i| / \max(|U|, |V_i|)$ , where  $|X|$  is the cardinality of the set  $X$ .
2. The best conserved operon pair  $(U, V^*)$  is the operon pair with the highest degree of conservation  $DC_i$ . This best operon pair is reported only if the degree of conservation is higher than the predefined threshold  $TDC$ ; otherwise, no conserved operon pair is returned.

The core of this algorithm is to calculate *mwmcm*. A *matching* in a graph  $G = (V, E)$  is a subset  $M$  of the edges  $E$  such that no edges in  $M$  share a common vertex, and a *maximum cardinality matching* (*mcm*) is a matching with the highest possible cardinality. An *mwmcm* is a *mcm* with the maximum total weight (see Figure 2 for an example). In this study, the edge relationship in an *mwmcm* represents the orthology relationship between the two corresponding operons. Using the scheme of *mwmcm* to identify the best conserved operon in *OPERMAP* has several advantages. First, it is guaranteed to find the maximum number of homologous gene relationships between two operons. Second, it can find the true orthologous gene pair based on sequence similarities in the case where there are several *mcms*, provided that an appropriate weighting scheme is given.

By applying *OPERMAP* on all reference genomes, we can obtain a set of conserved operons for a given query operon in the target genome. For each query operon out of 2,706 predicted operons in *E. coli*, we have applied *OPERMAP* on the 11 reference genomes. In this study, we want to cover not only fully conserved operons (*category-1* and *category-2*), but also partially conserved operons (*category-3*). Including partial conserved operons has its biological reasoning. Some large operons can break into multiple smaller operons with some part of these smaller operons still maintaining the same regulation mechanism. For instance, a Crp-regulated *xylFGHR* operon in *E. coli* breaks into *xylFGH* and *xylR* in *H. influenzae*, with *xylFGH* maintaining Crp regulation, but *xylR* not



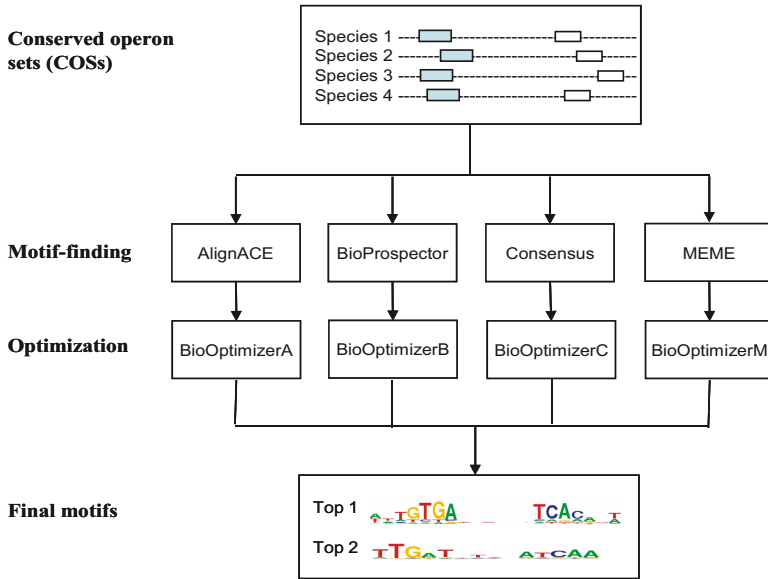
**Fig. 2.** An illustration of a maximum weight maximum cardinality matching (*mwmcM*). The resulting matching is shown on the right, with the matching size of 4. While the weight of the edge  $g_B - g_3$  is 20, the *mwmcM* does not choose it. Otherwise, the matching size will be 3.

[19]. Setting a low value of  $TDC$  (*i.e.*,  $< 0.5$ ) may introduce partial conserved operons with different regulation mechanisms. On the other hand, setting a high value of  $TDC$  (*i.e.*,  $> 0.8$ ) will exclude most of partial conserved operons with *category-3* since the sizes of most operons are less than five. We have chosen 0.6 for  $TDC$  in this study. Experiments on the determination of  $TDC$  will not be elaborated in this paper due to space limitation.

**Collection of regulatory sequences.** The gene annotations and the genomic sequences of the 12 genomes in this study were downloaded from the NCBI GenBank database (<http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi>). For each operon obtained in the previous step, we extract the upstream sequence up to 400 base-pairs (bp) from the translation start site, without overlap of the next upstream gene.

**Motif Discovery.** The upstream promoter sequences for each conserved operon are the input for our motif discovery pipeline to identify (possibly multiple) TFBSs. The pipeline is similar to our previously developed tool BEST [20], which contains four motif-finding programs: AlignACE [21], BioProspector [22], CONSENSUS [23] and MEME [24], as well as BioOptimizer [25] for optimizing the predictive power of each program. However, BEST is a graphic tool which makes it less suitable for the genome scale motif discovery. Our pipeline overcomes this drawback to produce top-ranked motifs for each sequence dataset in a fully automatic fashion. We outline our motif discovery pipeline in three stages (also see Figure 3).

1. Run the four motif-finding programs mentioned above. Since the motif length in all the four programs must be specified by the user, each program is run multiple times with different motif lengths ranging from 10 to 20 bp. The range of motif lengths chosen is based on the fact that most experimentally verified motifs fall in this range. For each width and each program, the top-ranked motif is collected, giving a set of  $4 \times 11 = 44$  top-ranked motifs.
2. The BioOptimizer program is run on each of the 44 motifs. BioOptimizer takes each predicted motif as the starting point and optimizes it using a local hill-climbing technique [25].
3. Rank all 44 optimized motifs based on their score values calculated by BioOptimizer, and output the top five.



**Fig. 3.** The workflow of motif discovery. Upstream sequences of conserved operons among closely related species are generated by *OPERMAP*. Datasets are fed into multiple motif-finding programs, and candidate discovered motifs are then optimized by BioOptimizer. The top-ranked motifs based on the score function of BioOptimizer are final identified motifs.

**Performance Evaluation.** We validate our motif predictions with a similar approach to past motif discovery investigations. We define as *true positives* ( $TP$ ) the predicted binding sites which overlap with the true binding sites by at least 50%; *false positives* ( $FP$ ) are the predicted binding sites which have no such overlap; *false negatives* ( $FN$ ) are the true binding sites that have no overlap with any of the predicted binding sites. We focus on four validation measures, sensitivity ( $Sn$ ), specificity ( $Sp$ ), performance coefficient ( $PC$ ), and F-measure ( $F$ ), which are defined as follows:

$$Sn = TP / (TP + FN) \quad (1)$$

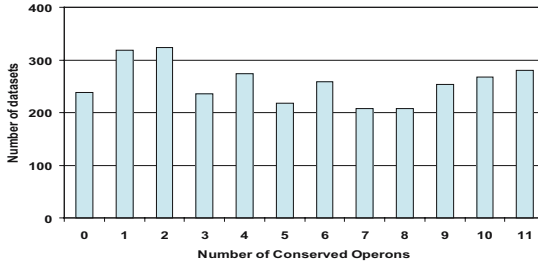
$$Sp = TP / (TP + FP) \quad (2)$$

$$PC = TP / (TP + FN + FP) \quad (3)$$

$$F = 2 * Sn * Sp / (Sn + Sp) \quad (4)$$

### 3 Results

**Collection of Conserved Operons.** The genome sizes of our 12 genomes range from 4.2 Mb to 6.9 Mb, and the numbers of predicted operons ranged from 1596 to 4468. For each of the 2,706 predicted operons in *E. coli*, we ran



**Fig. 4.** The operon conservation histogram for 2706 predicted operons of *E. coli*. X-axis indicates the number of conserved operons in 11 other species, and y-axis indicates the number of conserved operons with the conservation number ranging from 0 to 11.

*OPERMAP* to identify conserved operons in the 11 reference genomes. The distribution of the number of conserved operons across the twelve genomes is shown in Figure 4. Two hundred and thirty-eight operons (8.8%) from *E. coli* do not have a corresponding operon match in any of the 11 reference genomes, which may indicate that those operons are either unique to *E. coli* or have been predicted incorrectly by UNIPOPOP. At the opposite extreme, 280 operons (10.3%) are conserved across all 11 reference genomes.

**Performance of TFBS Predictions.** Our evaluation was restricted to predicted motifs in conserved operon sets in *E. coli* since experimentally-verified binding-sites are not available in the 11 reference genomes. We retrieved verified binding sites of *E. coli*, grouped by transcription factors, from the PRODORIC database [26]. We focus on the binding sites regulated by the following ten transcription factors: ArgR, Crp, Fis, Fnr, Fur, IHF, LexA, Lrp, MetJ and SoxS, totally covering 424 verified binding sites. Table 1 shows individual performance statistics for each transcription factor. Prediction accuracies vary among 10 TFs. For example, the prediction sensitivity was 92.6% for LexA, but only 46.7% for Lrp with the known motif. Further studies have shown that Lrp-associated motif was quite degenerate, with the pattern of “NNNNNNTTTTATTCT”, thus making motif-finding quite difficult. In contrast, LexA-associated motif was a 16-bp palindrome, with a conserved pattern of “CTGTATATATATACAG”. In general, our motif discovery pipeline has a high sensitivity but low specificity, similar to other motif prediction results [17]. However, some of this low specificity could be due to unverified but true sites. As more binding sites are verified and deposited in the PRODORIC database, some predicted false positives could become true positives.

**Comparison to other approaches.** We also compared the performance of our conserved operon-based approach with two orthologous gene-based (specifically sequence similarity-based) approaches, which were used in MicroFootprinter [15] and PHYLOCLUS [16] respectively. In both methods, orthologous genes in other species were identified using a reciprocal BLAST best-hit procedure, with a

**Table 1.** Prediction accuracy of motif-findings on 10 TFBSs of *E. coli* using the PFP approach

TFs	ArgR	Crp	Fis	Fnr	Fur	IHF	LexA	Lrp	MetJ	SoxS
<i>Sn</i>	0.682	0.64	0.5	0.655	0.761	0.5	0.926	0.467	0.818	0.722
<i>Sp</i>	0.205	0.094	0.113	0.113	0.181	0.066	0.116	0.109	0.138	0.088
<i>PC</i>	0.188	0.089	0.102	0.107	0.172	0.061	0.115	0.097	0.134	0.086
<i>F</i>	0.316	0.163	0.185	0.193	0.293	0.116	0.206	0.177	0.237	0.158

**Table 2.** Performance comparison between the conserved operon-based (PFP) and the orthologous gene based approaches. The one used in ‘Microfootprinter’ is named as OrthM, while the one used in ‘PHYLOCLUS’ is named as OrthB.

Methods	<i>Sn</i>	<i>Sp</i>	<i>PC</i>	<i>F</i>
OrthM	0.605	0.109	0.102	0.184
OrthB	0.603	0.105	0.098	0.179
PFP	0.636	0.106	0.100	0.182

**Table 3.** A list of *glnHPQ* associated orthologous genes and conserved operons predicted by OrthM, OrthB and PFP. *glnH* from *E. coli* was used as a query gene in OrthM and OrthB, while *glnHPQ* from *E. coli* was used as a query operon in PFP. The degree of operon conservation was calculated by *OPERMAP*.

Species	OrthM	OrthB	PFP	Degree of Conservation
<i>E. coli</i>				
<i>A. hydrophila</i>				0.67
<i>E. carotovora</i>				1
<i>P. profundum</i>				
<i>P. luminescens</i>				
<i>P. fluorescens</i>				1
<i>S. enterica</i>				1
<i>S. ANA</i>				1
<i>S. sonnei</i>				1
<i>V. parahaemolyticus</i>				
<i>Y. pestis</i>				1

threshold of  $10^{-6}$ . For each method, we generated sequence data sets, ran our motif pipeline for TFBSs prediction, and then evaluated predictions based on 424 binding sites from the PRODORIC database. As shown in Table 2, our approach was more sensitive than the two other ones (63.6% versus 60.5% and 60.3%). The higher sensitivity of our approach over the other two can be attributed to

the reliability of our generated orthologous promoter regions. For example, our approach could detect the true binding-sites of the glutamine permease operon *glnHPQ* in *E. coli*, while the orthologous gene-based couldn't. An investigation of the datasets showed that our approach identified 7 conserved operons for *glnHPQ*, while 'OrthM' identified 10, and 'OrthB' identified 6 "orthologous" genes for *glnH* (shown in Table 3). Further analysis has shown that three 'orthologous' genes (e.g., 117619357, *artI*, 2800492) found by 'orthM' were actually arginine ABC transporters. In addition, both 'orthB' and 'orthM' considered '70728423' from *P. fluorescens* to be an 'orthologous' gene for *glnH*, while our approach did detect a conserved operon *glnHP-70733921*. All these indicate that these four identified genes are not true orthologues, and introduction of the sequences of these genes in OrthB and OrthM lead to the reduction of information content for motif finding.

## 4 Conclusion

We have presented a computational framework of phylogenetic footprinting in prokaryotes. The major contributions of our work include: a) the introduction of the conserved operon approach, rather than the orthologous gene approach, to collect promoter sequence datasets, and b) the development of motif-discovery pipeline for identifying TFBSs from the sequences we have identified. Performance comparison of TFBSs prediction between our approach and others has shown that our approach could identify more experimentally verified binding-sites.

The better performance of our approach over previous ones is mainly due to the followings: the correct characterization of operon structures in the recent research efforts, and the correct determination of orthology relationships by relying on multiple homologous gene relationships within an operon. In addition, our algorithm *OPERMAP* can nicely incorporate three different categories of conserved operons that maintain the same regulation mechanism.

In our future work, we will predict TFBSs of prokaryotes at the genome scale using our computational framework. By clustering these predicted TFBSs, we can ultimately decipher regulons, which is the set of operons whose promoter regions share the similar binding motif patterns regulated by the same transcription factor.

**Acknowledgments.** This research was supported in part by National Science Foundation (#NSF/DBI-0354771, #NSF/ITR-IIS-0407204, #NSF/DBI-0542119, and #NSF/CCF-0621700) and by a "distinguished scholar" grant from Georgia Cancer Coalition.

## References

1. Tagle, D.A., Koop, B.F., Goodman, M., Slightom, J.L., Hess, D.L., Jones, R.T.: Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crasicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints 203, 439–455 (1988)
2. Tatusov, R.L., Koonin, E.V., Lipman, D.J.: A genomic perspective on protein families. *Science* 278, 631–637 (1997)

3. Li, L., Stoekert Jr, C.J., Roos, D.S.: OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome research* 13, 2178–2189 (2003)
4. Wang, T., Stormo, G.D.: Identifying the conserved network of cis-regulatory sites of a eukaryotic genome. In: *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, pp. 17400–17405 (2005)
5. Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., Kellis, M.: Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* 434, 338–345 (2005)
6. Wu, H., Mao, F., Olman, V., Xu, Y.: Accurate prediction of orthologous gene groups in microbes. In: *Proceedings/ IEEE Computational Systems Bioinformatics Conference, CSB*, pp. 73–79 (2005)
7. Westover, B.P., Buhler, J.D., Sonnenburg, J.L., Gordon, J.I.: Operon prediction without a training set. *Bioinformatics (Oxford, England)* 21, 880–888 (2005)
8. Ermolaeva, M.D., White, O., Salzberg, S.L.: Prediction of operons in microbial genomes. *Nucleic acids research* 29, 1216–1221 (2001)
9. Che, D., Zhao, J., Cai, L., Xu, Y.: Operon Prediction in Microbial Genomes Using Decision Tree Approach. In: *Proceedings of IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pp. 135–142 (2007)
10. Price, M.N., Huang, K.H., Alm, E.J., Arkin, A.P.: A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic acids research* 33, 880–892 (2005)
11. Dam, P., Olman, V., Harris, K., Su, Z., Xu, Y.: Operon prediction using both genome-specific and general genomic information. *Nucleic acids research* 35, 288–298 (2007)
12. McCue, L., Thompson, W., Carmack, C., Ryan, M.P., Liu, J.S., Derbyshire, V., Lawrence, C.E.: Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic acids research* 29, 774–782 (2001)
13. McCue, L.A., Thompson, W., Carmack, C.S., Lawrence, C.E.: Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome research* 12, 1523–1532 (2002)
14. McGuire, A.M., Hughes, J.D., Church, G.M.: Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome research* 10, 744–757 (2000)
15. Neph, S., Tompa, M.: MicroFootPrinter: a tool for phylogenetic footprinting in prokaryotic genomes. *Nucleic acids research* 34, 366–368 (2006)
16. Jensen, S.T., Shen, L., Liu, J.S.: Combining phylogenetic motif discovery and motif clustering to predict co-regulated genes. *Bioinformatics (Oxford, England)* 21, 3832–3839 (2005)
17. Hu, J., Li, B., Kihara, D.: Limitations and potentials of current motif discovery algorithms. *Nucleic acids research* 33, 4899–4913 (2005)
18. Mehlhorn, K., Näher, S.: *Leda: a platform for combinatorial and geometric computing*. Cambridge University Press, Cambridge (1999)
19. Tan, K., Moreno-Hagelsieb, G., Collado-Vides, J., Stormo, G.D.: A comparative genomics approach to prediction of new members of regulons. *Genome research* 11, 566–584 (2001)
20. Che, D., Jensen, S., Cai, L., Liu, J.S.: BEST: binding-site estimation suite of tools. *Bioinformatics (Oxford, England)* 21, 2909–2911 (2005)
21. Roth, F.P., Hughes, J.D., Estep, P.W., Church, G.M.: Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nature biotechnology* 16, 939–945 (1998)

22. Liu, X., Brutlag, D., Liu, J.: BioProspector: discovering conserved DNA motifs in upstream regulatory regions of coexpressed genes. *Pac. Symp. Biocomput.*, 127–138 (2001)
23. Hertz, G.Z., Stormo, G.D.: Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics (Oxford, England)* 15, 563–577 (1999)
24. Bailey, T.L., Elkan, C.: Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28–36. AAAI Press, Menlo Park, California (1994)
25. Jensen, S.T., Liu, J.S.: BioOptimizer: a Bayesian scoring function approach to motif discovery. *Bioinformatics (Oxford, England)* 20, 1557–1564 (2004)
26. Munch, R., Hiller, K., Grote, A., Scheer, M., Klein, J., Schobert, M., Jahn, D.: Virtual Footprint and PRODORIC: an integrative framework for regulon prediction in prokaryotes. *Bioinformatics (Oxford, England)* 21, 4187–4189 (2005)