Computational Prediction of Operons in *Synechococcus sp.* WH8102

Xin Chen[1] xinchen@cs.ucr.edu
Zhengchang Su[2] zhx@csbl.bmb.uga.edu
Ying Xu[2] xyn@bmb.uga.edu
Tao Jiang[1] jiang@cs.ucr.edu

[1] Department of Computer Science and Engineering, University of California at Riverside, CA 92507, USA
[2] Department of Biochemistry and Molecular Biology, University of Georgia at Athens, GA 30602, and Computational Biology Institute, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

Abstract

We computationally predict operons in the *Synechococcus sp.* WH8102 genome based on three types of genomic data: intergenic distances, COG gene functions and phylogenetic profiles. In the proposed method, we first estimate a log-likelihood distribution for each type of genomic data, and then fuse these distribution information by a perceptron to discriminate pairs of genes within operons (WO pairs) from those across transcription unit borders (TUB pairs). Computational experiments demonstrated that WO pairs tend to have shorter intergenic distances, a higher probability being in the same COG functional categories and more similar phylogenetic profiles than TUB pairs, indicating their powerful capabilities for operon prediction. By testing the method on 236 known operons of *Escherichia coli* K12, an overall accuracy of 83.8% is obtained by joint learning from multiple types of genomic data, whereas individual information source yields accuracies of 80.4%, 74.4%, and 70.6% respectively.

We have applied this new approach, in conjunction with our previous comparative genome analysis-based approach, to predict 556 (putative) operons in WH8102. All predicted data are available at (`http://www.cs.ucr.edu/~xin/operons.htm`) for public use.

Keywords: operon, log-likelihood, intergenic distance, COG function, phylogenetic profile

1 Introduction

Operons represent a basic organizational unit of genes in the complex hierarchical structure of biological processes in a cell of prokaryotes. They are mainly used to facilitate efficient implementation of transcriptional regulation in microbial genomes [24]. Operons provide highly useful information for the characterization and construction of biological pathways and networks at a large scale. Therefore, the prediction of operons at the whole-genome level is one of the most fundamental and challenging computational problems in microbial functional genomics.

A great amount of research effort has been devoted in the past several years to the investigation of effective methods for operon prediction, and a number of algorithms have been developed [4, 5, 7, 8, 11, 19, 20, 25, 26]. Through these studies several observations have been made: (i) Intergenic distances are highly effective in discriminating WO pairs from TUB pairs, among most if not all prokaryotes, and (ii) Joint learning from heterogeneous characterization information of the genome significantly improve operon prediction capability. Besides, most of these methods require many experimental data as input so that they are generally not applicable to newly sequenced genomes.

Recent methodological advances enable us to computationally collect many different types of data to help the prediction of operon structures in microbial genomes. These data include genes (*e.g.* by Glimmer [3]), their functional categories (*e.g.*, by COG [22]), promoters (*e.g.*, by SIGSCAN [15]), terminators (*e.g.*, by TransTerm [6]), and phylogenetic profiles (*e.g.*, by BLASTp [1]). However, they

could be erroneous largely due to our current limited understanding of these biological mechinary. Errors may prevent a single type of data from predicting operon successfully as it was expected. However, if we assume that errors across different genomic data types are largely independent then joint learning from them could lead to new insights that might not be as readily available from analyzing one type of data in isolation, and thus increase the operon prediction accuracies significantly.

Our goal is to devise a compuational method for operon prediction, that can be directly applicable to recently sequenced genomes that have not been under extensive experimental investigation. This is very important in practice because putative operons could provide biologists with first insights as soon as a genome is sequenced and its genes are annotated. As our first attempt towards this goal, a computational method based on comparative genome analysis has been proposed [5], which searches for operons conserved in closely related species. Definitely, this method will miss operons that are unique to the species of interest and not conserved in other species. In order to overcome this problem, we present in this paper a new compuational method for operon prediction, through fusing information from three different types of genomic data: intergenic distances, COG gene functions and phylogenetic profiles. All these genomic data can be computationally obtained, which is an apparent advantage over many existing methods that make operon prediction by joint learning from multiple type of experimental genomic data, for example, gene microarray expression data [4, 20], gene function annotations [19] and metabolic pathways [26].

In the paper, we first provide evidence that intergenic distances, gene functional categories assigned by COG and phylogenetic profiles calculated from 145 microbial genomes work surprisingly very well for discriminating WO pairs from TUB pairs. Because the promoters and terminators predicted by using SIGSCAN [15] and TransTerm [6], respectively, could not provide sufficient discriminating power [5], they were left out of our genomic dataset for operon prediction. Compared to TUB pairs, WO pairs tend to have shorter intergenic distances, a higher probability of being in the same COG functional categories, and having more similar phylogenetic profiles (*i.e.*, co-evolving pattern). Second, we estimate a log-likelihood distribution for each type of genomic data using 236 known operons that have been experiementally verified in *Escherichia coli*. Third, we devise a "knowledge fusion" based prediction algorithm, which is a *perceptron* using log-likelihood scores as input. It makes prediction by fusing knowledge extracted from various sources of genomic data. Tested on the known operons, the new method consistently outperforms any method which makes prediction by learning from only a single type of genomic data, in terms of sensitivity, specificity and accuracy (See definition below). For example, the overall accuracy increases up to 83.8% by joint prediction from 80.4%, 74.4% and 70.6% by prediction based on only one type of genomic data. Finally, we apply the new computational method to the *Synechococcus sp.* WH8102 genome and obtain 917 WO pairs, resulting in 537 putative operons.

In addition, we developed an operon database for the *Synechococcus sp.* WH8102 genome where operons came from the prediction by the computational method we proposed here or from the prediction by the comparative genome analysis-based approach in our previous effort [5]. Various experimental evidence that support our prediction, *e.g.*, COG function categories and phylogenetic profiles, can also be found in the database. This operon database will be very useful to research related to the *Synechococcus sp.* WH8102 genome.

2 Methods

2.1 Data Preparation

All the genome sequences and their annotated genes are downloaded from the GenBank database (`http://www.ncbi.nlm.nih.gov/genomes/MICROBES/Complete.html`). There are totally 145 microbial organisms that were completely sequenced (as of January 11, 2004), and selected for the phylogenetic profiling construction. They are also listed on our website (`http://www.cs.ucr.edu/~xinchen/operons.htm`). We choose two specific microbes, *Escherichia coli* K12 (GenBank accession number

NC_000913) and *Synechococcus sp.* WH8102 (GenBank accession number NC_005070), to present our experimental results.

In the RegulonDB database [21], there are 237 known operons of *E. coli* that are experimentally verified. However, because there are some genes that have been removed from the latest annotation of *E. coli*, only 236 operons are actually left for our experiments. These operons include 568 WO pairs (*i.e.*, adjacent genes within operons, and 463 TUB pairs (*i.e.*, adjacent genes across the borders of operons).

Since two genes of a WO pair must be on the same DNA strand with no intervening gene in the complementary strand, we evaluate only adjacent pairs having the same transcritional direction in our predictive algorithm, and two adjacent genes will be simply classified as a TUB pair if they are in the opposite strands of a genome.

2.2 Performance Measurement

A common way to measure the performance of a predictive approach is to estimate an accuracy score from a test dataset. The accuracy, which follows the definition given in [11], is the average of sensitivity and specificity, where sensitivity is the fraction of true positives detected in the total known WO pairs (568 pairs in our experiments) and specificity is the fraction of true negatives in the total known TUB pairs (463 pairs).

2.3 Log-Likelihoods

What we can observe usually is the properties of an object, but we may not know exactly what it is. Therefore, a typical prediction problem is how to identify an object given its observed properties. For example, an operon prediction problem is to determine whether two adjacent genes belong to a same operon if we only know their genomic sequence locations. Using probability measurements, this can then be formulated as a problem to calculate a posterior probability, $P(\mathrm{WO}|(g_a, g_b))$, where g_a and g_b are the property values observed for two adjacent genes of interest. By Bayesian theorem, we have

$$P(\mathrm{WO}|(g_a, g_b)) = \frac{P((g_a, g_b)|\mathrm{WO})}{P((g_a, g_b))} P(\mathrm{WO})$$

where $P((g_a, g_b)|\mathrm{WO})$ is the probability that a specific property value (g_a, g_b) could be observed given a WO gene pair, $P((g_a, g_b))$ is an unconditional version of $P((g_a, g_b)|\mathrm{WO})$, *i.e.*, the marginal probability that a specific property value could be observed from a randomly selected gene pair which may or may not be in an same operon, and $P(\mathrm{WO})$ is the probability that a gene pair is expected to be in an operon without any prior information, which in general is set to be constant in a typical prediction problem.

In most real applications, however, it seems highly problematic to make an accurate statistical measurement for $P(\mathrm{WO}|(g_a, g_b))$ because we may observe a huge amount of property instances while only a small amount of training data is available. To deal with this problem, we first try to define a distance measure between two property values $d(g_a, g_b)$ (*e.g.*, intergenic distance between two genes or Hamming distance between two binary vectors of phylogenetic profiles) and then calculate $P(\mathrm{WO}|d(g_a, g_b))$ instead of $P(\mathrm{WO}|(g_a, g_b))$. This may dramatically reduce the number of instances for the estimation of posterior probability. Actually, $P(\mathrm{WO}|d(g_a, g_b))$ could be equal to $P(\mathrm{WO}|(g_a, g_b))$ if we assume any pair of (g_a, g_b) at a same distance has the same probability being a true WO pair. In order to achieve a good prediction performance, a distance measure to be defined should reflect a simple intuition that a shorter distance implies a higher probability of a gene pair belonging to the same operon.

For most typical prediction problems, there are only two exclusive outputs, *i.e.*, the input object is or is not the one we are interested in. For example, a pair of genes is either a WO pair or a TUB pair exclusively. It thus follows that

$$P(d(g_a, g_b)) = P(d(g_a, g_b)|\mathrm{WO})P(\mathrm{WO}) + P(d(g_a, g_b)|\mathrm{TUB})P(\mathrm{TUB})$$

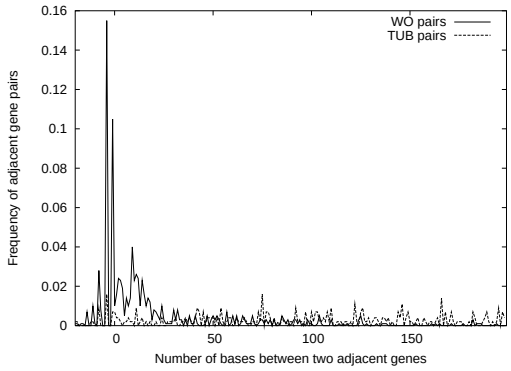

Figure 1: Frequency distribution of intergenetic distances between adjacent gene pairs.

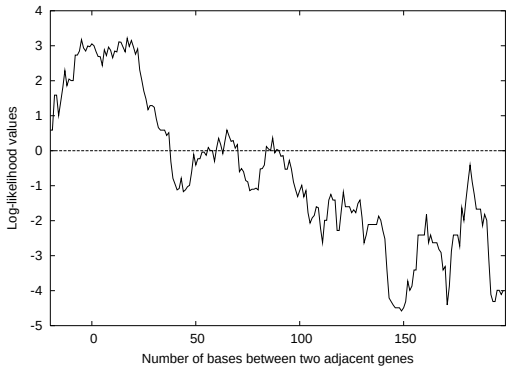

Figure 2: Log-likelihoods for intergentic distances between adjacent gene pairs.

Assuming that $P(\text{WO})$ is a constant (as it is in many real applications), we can see that the ratio of $P(d(g_a, g_b)|\text{WO})$ against $P(d(g_a, g_b)|\text{TUB})$ reflects the same relationship among gene pairs at different distances as the posterior probability value of $P(\text{WO}|d(g_a, g_b))$ does. In other words, the ratio strictly increases as the probability increases. This ratio is commonly called a likelihood score, and a logarithm likelihood score is then defined as follows:

$$LL(\text{WO}|d(g_a, g_b)) = log\frac{P(d(g_a, g_b)|\text{WO})}{P(d(g_a, g_b)|\text{TUB})}$$

Therefore, the higher a log-likelihood score is the more likely two genes form a WO pair.

2.4 Log-Likelihoods from Intergenic Distances

The distance between two adjacent genes on the same strand of DNA tends to be shorter if they belong to the same operon. This characterization has been observed by many researchers [19, 11]. Experimental evidence was also given that this method is applicable to most if not all prokaryotic genomes, indicating short distance tendency between adjacent pairs is one of universal structure features of operons.

The intergenetic distance is calculated as the number of bases between the two closest end points of two genes in the genomic sequences. We estimated intergenetic distance frequencies of known WO pairs and of known TUB pairs, and then calculated their log-likelihood scores using the formula described above. Experimental results are depicted in Fig. 1 which plots the intergenic distance frequencies, and in Fig. 2 which plots their log-likelihood scores. As demonstrated in [21], the two most frequent distances are $-4bp$ and $-1bp$ (*i.e.*, two genes overlap by 4 bases or 1 base).

2.5 Log-Likelihoods from COG Functions

Genes in an operon are involved in a specific functional process or pathway. These genes, therefore, are expected to share the same functional category. Salgado *et al.* [19] has demonstrated this by analyzing the functional classes of Monica Riley [18], which were experimentally annotated for the genes examined. Instead, we took a computational method where a putative gene was assigned a COG function category by the COGnitor program from `http://www.ncbi.nlm.nih.gov/COG/`, and then estimated its discrimination power to WO pairs against TUB pairs.

The COGs (*i.e.*, clusters of orthologous groups, each of which consists of individual orthologous genes or orthologous sets of paralogs from at least three lineages) comprise a framework for the analysis of evolutionary and functional relationships among homologous genes from multiple genomes [22]. There are three major levels in the COG function hierarchy. The first level, for example, consists of four categories: (1) Information storage and processing, (2) Cellular processes, (3) Metabolism, and (4)

Poorly characterized. Because COGnitor may fail to assign some new genes to a functional category or to the poorly characterized category, pairs involving such genes will be ignored in the subsequent statistics calculation and simply assigned a likelihood score of zero. As our previous experiments have shown [5], two genes of a WO pair are more likely to share the same COG functional category than those of a TUB pair. For example, there are a total of 79.5% (272 out of 342) of the known WO pairs sharing the same COG second-level functional category. As a comparison, only 31.0% (50 out of 161) of the known TUB pairs share the same second-level category.

We further found that the probability that two adjacent genes share a same COG second-level functional category varies as different first-level functional categories are examined. This then results in unequal log-likelihood scores calculated for adjacent gene pairs in the different first-level functional categories, as shown in Table 1. In particular, two genes that are predicted in the COG functional category *Metabolism* are more likely to be a WO pair than those in the category *Information storage and processing* or *Cellular processes*.

Table 1: Frequencies of adjacent pairs in COG functional categories and their log-likelihood scores.

COG functional categories	% WO pairs	% TUB pairs	Log-likelihoods
Information storage and processing	0.143	0.074	0.942
Cellular processes	0.187	0.093	1.006
Metabolism	0.464	0.142	1.702
Different characterized categories	0.204	0.689	-1.75

2.6 Log-Likelihoods from Phylogenetic Profiles

2.6.1 Phylogenetic Profiles

The phylogenetic profile [16] of a gene is a binary string, each bit of which represents the presence or absence of the gene in the genome of an organism. Phylogenetic profiles have been demonstrated in various applications to be a powerful source of information for assigning gene function and inferring protein interaction in comparative genomics [9, 10, 12, 13]. Its basic assumption from the biological aspect is that proteins functioning together in a pathway or structural complex are likely to evolve in a correlated fashion. Compared to traditional homology-based methods, an apparent advantage of the new method is that pairs of functionally linked genes inferred do not necessarily have amino acid sequence similarity with each other.

Instead of a binary value representing the homologous presence or absence of a protein in an organism, Marcotte [12] proposed to use a fractional number ranging from zero to one as an entry of the vector representing a phylogenetic profile. This number, calculated based on E-values from BLASTp program, reflects the sequence similarity level between a pair of homologous genes. The advantage of this formulation over the original one is unknown, because no experimental comparison between them has been reported in the literature.

A variant of phylogenetic profiles is *inverse phylogenetic profiles*, *i.e.*, a profile where ones have been replaced by zeros and vice versa [10]. A *convergent evolution* is expected to be found in genes that perform redundant functions but lack a common evolutionary origin. This will involve genes that perform the same biological function, but are not homologs or may not even perform the same biochemical function. Such genes could possibly be detected by computing an inverse phylogenetic profile with a phylogenetic profile. In general, the existence of an inverse profile that is the same or similar to a profile should imply either functional similarity or alternatively mutually exclusive functionalities [10].

2.6.2 Previous Distance Measurements

A basic issue arising in any computational analysis of phylogenetic profiles is how to measure the distance between two profiles. In principle, distance values to be calculated should reflect functional

relevance of two proteins as more accurately as possible. Undoubtedly, this is not a trivial job. In the following, we briefly summarize several distance measurements we found in the literature.

Hamming distance: The Hamming distance was proposed as early as the concept of phylogenetic profile was first introduced [16]. It simply counts the number of bits that differ between two profile strings.

Differential parsimony: The next distance measure, which was introduced in [10], calculates a differential parsimony in the historical evolution of two genes based on their phylogenetic profiles. It requires the phylogenetic reconstruction of the ancestral proteins and the comparison of the reconsructed trees. Compared to Hamming distance, this distance measure is evolutionarily relevant, and thus more biologically meaningful.

Tree kernel: With a powerful mathematical framework embedded, a tree kernel is proposed to analyze phylogenetic profiles [23]. By defining a tree kernel, *i.e.*, an inner product function in the high-dimensional feature space mapped from the phylogenetic profiles, a distance measure is then defined in an implicit way. Kernel methods work implicitly in the feature space using only the kernel function, for which existing diverse algorithms such as Support Vector Machines for classification or regression, kernel principle component analysis, kernel clustering or kernel Fisher discriminants can directly be applied to the defined tree kernel [23].

2.6.3 Entropy-Based Distance

We observe that the distribution pattern of 0-1 identities between two phylogenetic profiles could make a difference in the inference of the co-evolving relationship of two genes. For example, there are many genes that exist in all organisms or may not exist in any organism examined, and their profiles are all ones or zeros, respectively. Intuitively, these genes are impossible to be functionally related, but a small distance value would be obtained if we use any of the above distance definitions (*i.e.*, Hamming distance, differential parsimony or the tree kernel). If the bit identities involving 0's and 1's between two profiles are normally distributed over all the organisms then we could perhaps expect more on a co-evolving pattern and functional association.

In order to address this issue, the new distance measure proposed in this paper is an extension of Hamming distance by taking into account the distribution pattern of 0-1 identities between two profiles. Since identities involving only 0's or only 1's suggest the least co-evolving possibilities between two genes but that involving the same number of 0's and 1's suggest the highest, we propose a new distance measure d_E as follows:

$$d_E = L - (L - d_H)\sqrt{E(p)}$$

where d_H is the Hamming distance between two profiles, L is the length of a binary string profile (*i.e.*, the number of genomes examined), and p is the percentage of identities involving 0's among in all the identities and $E(p)$ is its Shannon entropy, *i.e.*, $E(p) = -plog(p) - (1-p)log(1-p)$. We can see that, d_E becomes the Hamming distance measure when $E(p)$ is set to be 1, which means that an equal weight is assigned to any distribution pattern of 0-1 identities. The square root is employed here in order to widen the curve plotted by the entropy function $E(p)$, and to produce a slightly higher prediction accuracy than using the measure without a square root.

We conducted experiments to compare the entropy-based distance with Hamming distance. The entropy-based distance performs superior to Hamming distance when applied to operon prediction, in terms of prediction accuracies estimated on known WO pairs and known TUB pairs. Figure 3 shows that the new entropy-based distance can increase the maximal accuracy from 0.670 (as given by Hamming distance) to 0.713. As we mentioned earlier, prediction accuracy reflects a distance measure's capability to discriminate WO pairs from TUB pairs. The frequency distributions of Hamming distance and the entropy-based distance for adjacent (WO or TUB) pairs of genes, as well as their log-likelihood scores, are depicted in Figure 4.

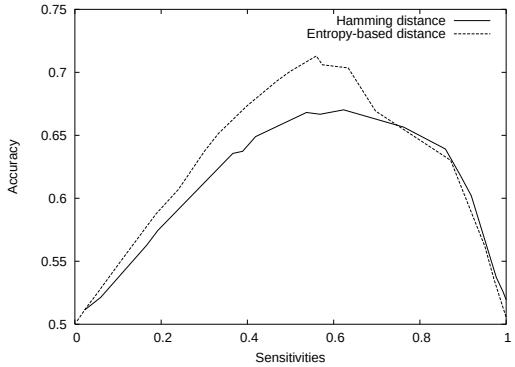

Figure 3: Accuracy comparison at different sensitivities between Hamming distance and entropy-based distance.

2.7 Predicting Operons from Multiple Types of Data

In many typical prediction problems in bioinformatics, what we could know about an object is various disparate types of property information associated with it. It may be easy to make a prediction based on only one type of data. However, a problem is arising if we want to make a prediction based on a heterogeneous data set, *i.e.*, how to integrate disparate types of information data so as to improve the prediction capability. For example, we can make operon prediction by joint learning from three types of characterization information described above – intergenic distances, COG functional categories and phylogenetic profiles. They are completely disparate because an intergenic distance is measured by the number of nucleitide bases separating two genes (which could be negative if the genomic sequences of two genes overlap), a COG functional category is labelled by a symbol, and phylogenetic profiles are encoded as binary vectors.

Prediction/classification based on multiple types of information actually is a research focus in the field of information fushion, which mainly develops techniques exploiting information dependences for better prediction performance. We choose a simple but powerful tool, *perceptron*, to implement our operon prediction task.

2.7.1 Perceptron

A perceptron is a single-layer network with threshold activation functions [2]. The perceptron we use for operon prediction has three input values, each corresponding to a log-likelihood score calculated from a type of genomic data, and its linear discriminant function is

$$y(LL) = w^T \cdot LL + b$$

where w is the weight vector, b is the bias (its negative is sometimes called a threshold) and LL is a vector with three components, *i.e.*, LL_i, $i = 1, 2, 3$. The output unit activation function is most conveniently chosen to be an anti-symmetric version of the threshold activation function of the form

$$g(a) = \begin{cases} -1 & \text{when } a < 0 \\ +1 & \text{when } a \geq 0 \end{cases}$$

That is, the two input genes are identified as a WO pair if $g(y(LL))$ is given $+1$, or a TUB pair if otherwise.

A perceptron must be trained to assign proper values to its weights and bias so as to produce an effective prediction system. In order to achieve this goal, it would be natural to define the error function in terms of the total number of misclassifications over a training data set, since the data set is usually not linearly separable in practice. Unfortunately, this error function seems very difficult to use in an efficient optimization algorithm, because it is a piecewise-constant function so that many optimization procedure akin to gradient descent cannot be applied [2].

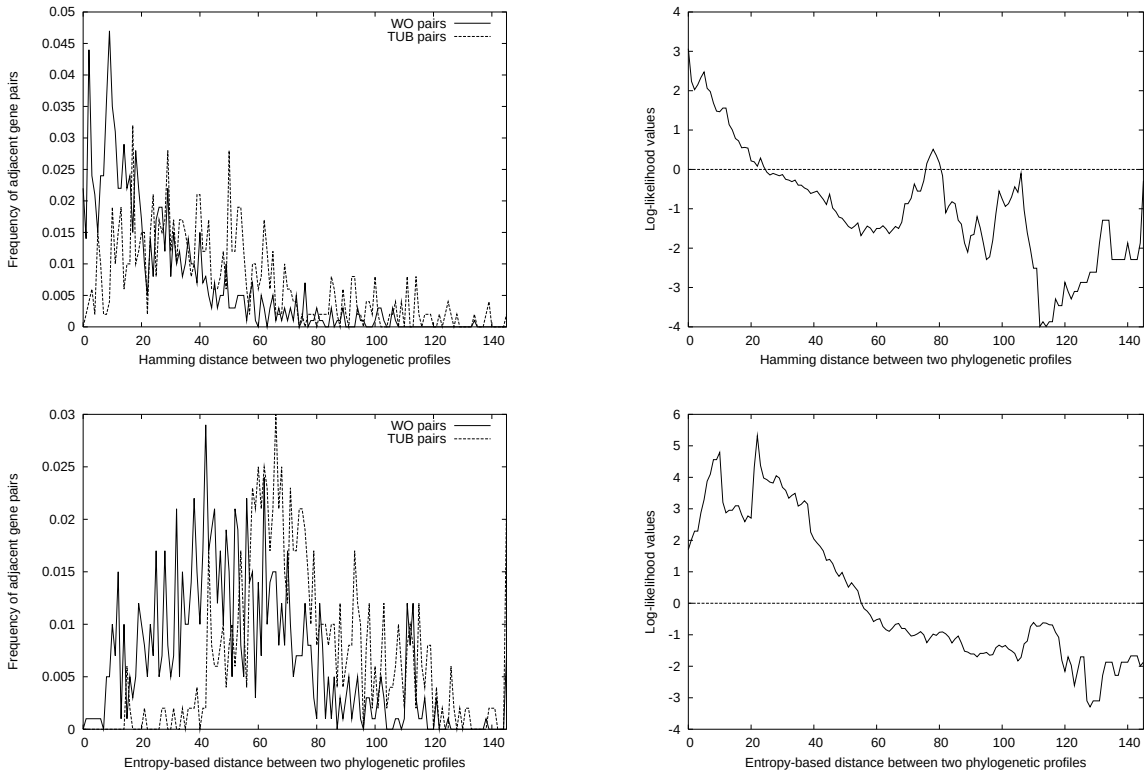

Figure 4: Frequency distribution and log-likelihood distribution of distances between two phylogenetic profiles of adjacent genes when Hamming distance and entropy-based distance are employed, respectively.

An alternative error function that is widely used is a continuous, piecewise-linear one called the perceptron criterion, which tries to minimize the Mean Absolute Error (mae)

$$E^{perc}(w) = \sum_{LL \in M} |y(LL)|$$

where M is the set of vectors LL that are misclassified by the current weight vector w. We can see that $E^{perc}(w)$ is proportional to the sum, over all of the input patterns that are misclassified, of the (absolute) distances to the decision boundary. The perceptron criterion is therefore continuous and piecewise linear with discontinuities in its gradient.

We have implemented the above perceptron in Matlab and chosen the Levenberg-Marquardt algorithm as its training function.

2.7.2 Diagram of Operon Prediction

A diagram of our operon prediction system is depicted in Figure 6. The prediction procedure begins with two genes of interest as input, and then computationally identifies each gene with three types of genomic data. In the next step, we estimate the distance between two adjacent genes for each data type, and calculate a corresponding log-likelihood score based on known operon data. Finally, three log-likelihood scores serve as inputs to a perceptron. We use known operons once more to train the perceptron to assign proper values to its weights and bias so that it could produce an effective operon prediction system.

A popular approach to joint learning from multiple types of data is the kernel method in which multiple types of data are directly input into a *support vector machine* learning algorithm [17]. In constrast, our operon prediction system employs a two-phase training procedure. The first phase that estimates log-likelihood scores corresponds to exploring information correlations within one type of data, whereas the second phase that estimates the perceptron weights and bias corresponds to exploring information correlation between different types. Therefore, the heterogeneous data is not

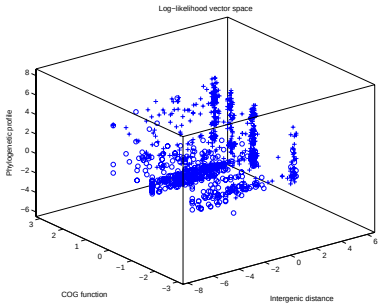

Figure 5: The log-likelihood feature space. WO pairs are plotted by plus sign, while TUB pairs by small circles.

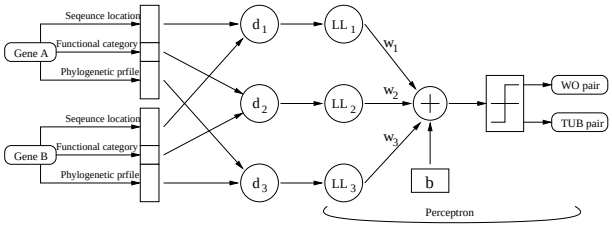

Figure 6: Diagram of our operon prediction system.

input to the perceptron directly. This could dramatically reduce the number of instances for the estimation of posterior probabilities. In addition, each training phase takes only one type of data as input, *e.g.*, log-likelihood scores in the second training phase.

3 Experimental Results

The log-likelihood feature vectors of 568 known WO pairs and 463 TUB pairs, each consisting of three log-likelihood scores estimated from intergenic distance, COG gene function and phylogenetic profile respectively, are plotted in Figure 5. We can see that the cluster of WO pairs separates very well with that of TUB pairs in the log-likelihood feature space, indicating the effectiveness of our method.

We use these feature vectors as input to train the perceptron, which explores prior knowledge of heterogeneity by assigning each data type a different weight. The resulted weights and bias are listed in Table 2, *i.e.*, $w^T = (1.0421\ 0.5138\ 0.5892)$ and $b = 0.1759$. We can see that intergenic distance was assigned a weight as almost twice as that of COG gene function or phylogenetic profile, showing that intergenic distance is the most importance feature among the three we examined.

Table 2: Weights and bias of the perceptron.

	Intergenic distance	COG function	Phylogenetic profile	bias
Weights	1.0421	0.5138	0.5892	0.1759

We run 10-fold cross-validation experiments with those known WO and TUB pairs, and estimate the sensitivities, specificities and accuracies of five operon prediction methods, including three methods based on each individual type of genomic data and two joint prediction methods with equal weights or weights assigned by training a perceptron. The results are shown in Table 3. Our prediction method employing a perceptron achieves the highest accuracy up to 83.8%, and consistently performs better than any prediction method based on only a single data type in terms of sensitivity, specificity and accuracy. The experimental results also show that log-likelihoods derived from intergenic distances contribute most of the discrimination power to the joint prediction method, as its assigned weight showed in the above. Compared with joint prediction with a perceptron, joint prediction using equal weights yields a slightly higher specificity, but it reduces the sensitivity much more so that its overall accuracy is still lower than that of the method employing a perceptron.

After the initial validation study, we applied our method to a newly sequenced microbial genome, *Synechococcus sp.* WH8102 [14], which is the focus of an ongoing US DoE GtL project (`http://www.genomes2life.org/`). This results in a total of 917 putative WO pairs, which then forms 537 putative operons consisting of 1454 genes in *Synechococcus sp.* WH8102. Note that the average size of these putative operons (*i.e.*, the average number of genes in each operon) is 2.71, which is very close to the statistics reported in [26].

Table 3: Sensitivity, specificity and accuracy of operon prediction.

Predictor	Sensitivity	Specificity	Accuracy
Intergenic distance	0.778	0.829	0.804
COG function	0.798	0.689	0.744
Phylogenetic profile	0.570	0.842	0.706
Joint prediction with equal weights	0.805	0.853	0.829
Joint prediction with a perceptron	0.824	0.851	0.838
Prediction by comparative analysis	0.829	0.721	0.775

Previously, we proposed a comparative genome analysis-based approach to predict operons in WH8102 that are conserved in closely related species, *i.e.*, *Prochlorococcus marinus* sp. MED4 and *Prochlorococcus marinus* sp. MIT9313 in our experiments [5], yielding an accuracy of 77.5%. It will miss operons unique to WH8102, while the new method presented here may miss some true operons with low log-likelihood scores. For example, the well-known carboxysome (SYNW1712, 1713, 1714, 1715, 1716, 1717, 1718, 1719) can be exactly found by the comparative approach but only partly by the new method. On the other hand, an ABC transporter operon (SYNW1915, 1916, 1917) can be exactly found by the new method but is missed by the comparative method. Therefore, we can see that these two computational methods can complement each other in operon prediction.

We further looked at 19 ABC transporter operons, a major family of transporters in WH8102. Their individual genes were identified as in [14] and collected manually into operons. Our new method was able to predict 10 of them exactly and did not miss any operon completely, better than the comparative approach which predicted only 8 operons exactly and missed 2, as shown in Table 4.

Table 4: The ABC transporter operons and our prediction results.

ABC transporter operon	Prediction by comparative genome analysis	Our prediction results
(SYNW0211, 0212)	Exactly found	Exactly found
(SYNW0319, 0320, 0321)	(SYNW0319, 0320, 0321, 0322)	(SYNW0318, 0319); (SYNW0320, 0321, 0322)
(SYNW0708, 0709)	Exactly found	(SYNW0708, 0709, 0710, 0711)
(SYNW0840,0841, 0842, 0843)	(SYNW0840,0841); (SYNW0842, 0843)	Exactly found
(SYNW0969, 0970)	Exactly found	Exactly found
(SYNW1086, 1087)	(SYNW1084, 1085, 1086, 1087)	(SYNW1084, 1085, 1086, 1087)
(SYNW1111, 1112)	(SYNW1109, 1110, 1111, 1112, 1113, 1114)	(SYNW1109, 1110, 1111, 1112, 1113)
(SYNW1168, 1169, 1170)	(SYNW1167, 1168); (SYNW1170, 1171)	(SYNW1166, 1167, 1168); (SYNW1169, 1170, 1171)
(SYNW1270, 1271, 1272)	Exactly found	Exactly found
(SYNW1283, 1284, 1285)	(SYNW1282, 1283, 1284, 1285)	(SYNW1281, 1282, 1283, 1284, 1285)
(SYNW1340, 1341)	Exactly found	Exactly found
(SYNW1415, 1416, 1417)	Not found	(SYNW1412, 1413, 1415, 1416, 1417, 1418)
(SYNW1797, 1798)	Not found	Exactly found
(SYNW1857, 1858)	(SYNW1857, 1858, 1859, 1860)	(SYNW1855, 1856, 1857, 1858, 1859, 1860)
(SYNW1915, 1916, 1917)	Exactly found	Exactly found
(SYNW2393, 2394, 2395)	Exactly found	(SYNW2393, 2394, 2395, 2396)
(SYNW2438, 2439, 2440, 2441, 2442)	(SYNW2438, 2439, 2440, 2441, 2442, 2443)	Exactly found
(SYNW2479, 2480, 2481)	(SYNW2479, 2480)	Exactly found
(SYNW2485, 2486, 2487)	Exactly found	Exactly found

In order to take advantege of the two computational methods, we incorporated all WO pairs predicted by either method together, resulting in 1108 WO pairs and 556 operons in WH8102. Among these WO pairs, 482 pairs were predicted by both methods, 191 pairs by the comparative method only and the other 435 pairs by the new method only. We have also developed an operon database for WH8102 where the user can find all putative operons and their supporting evidence such as COG gene functions and phylogenetic profiles. The database is at `http://www.cs.ucr.edu/~xinchen/operons.htm`.

4 Discussion

We compuatationally predicted operons in the *Synechococcus sp.* WH8102 genome based on three disparate types of genomic data: intergenic distance, COG gene function and phylogenetic profile. All these supporting data can be compuatationally obtained so that the proposed method is applicable to recently sequenced genomes that have not been under extensive experimental investigation. We demonstrated that WO pairs tend to have shorter intergenic distances, a higher probability of having

the same COG functional categories and more similar phylogenetic profiles than TUB pairs. Therefore, these structural features are very useful for the challenging operon prediction task.

We estimated a log-likehood distribution for each type of genomic data. For COG gene function, the log-likelihood scores vary as different first-level functional categories are examined. For phylogenetic profiles, we proposed an entropy-based distance by taking into account the distribution pattern of 0-1 identities, which is superior to Hamming distance when applied to operon prediction, in terms of prediction accuracies.

We devised an operon prediction system, which makes prediction by joint learning from multiple types of genomic data. It integrates two phases of training to explore information correlations within one type of data and information correlation between different types, respectively. We believe that this could be a technique with wide applications in the field of computational biology.

Acknowledgments

This research was supported in part by the US Department of Energy's Genomes to Life program (`http://doegenomestolife.org/`) under project, Carbon Sequestration in *Synechococcus sp.*: From Molecular Machines to Hierarchical Modeling (`http://www.genomes2life.org/`). YX's work is also supported, in part, by National Science Foundation (NSF/DBI-0354771, NSF/ITR-IIS-0407204). TJ's work is also partially supported by NSF grants ITR-0085910 and CCR-0309902, and National Key Project for Basic Research (973) grant 2002CB512801.

References

[1] Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D., Gapped BLAST and PSI-BLAST: A new generation of protein database search programs, *Nucleic Acids Res.*, 25:3389–3402, 1997.

[2] Bishop, C.M., *Neural Networks for Pattern Recognition*, Oxford University Press, 1995.

[3] Delcher, A.L., Harmon, D., Kasif, S., White, O., and Salzberg, S.L. Improved microbial gene identification with GLIMMER, *Nucleic Acids Res.*, 27:4636–4641, 1999.

[4] Craven, M., Page, D., Shavlik, J., Bockhorst, J., and Glasner, J., A probabilistic learning approach to whole-genome operon prediction, *Proc. 8th International Conference on Intellignet Systems for Mol. Biol.*, 116–127, 2000.

[5] Chen, X., Su, Z., Dam, P., Palenik, B., Xu, Y., and Jiang, T., Operon prediction by comparative genomics: An application to the *Synechococcus sp.* WH8102 genome, *Nucleic Acids Res.*, 32(7): 2147–2157, 2004.

[6] Ermolarva, M., Khalak, H., White, O., Smith, H. and Salzberg, S., Prediction of transcription terminators in bacterial genomes, *J. Mol. Biol.*, 301: 27–33, 2000.

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

[8] de Hoon, M.J.L., Imoto, S., Kobayashi, K., Ogasawara, N., and Miyano, S., Predicting the operon structure of *Bacillus subtilis* using operon length, intergenic distance, and gene expression information, *Proc. of the Pacific Symposium on Biocomputing*, 276–287, 2004.

[9] Huynen, M., Snel, B., Lathe, W., and Bork, P., Predicting protein function by genomic context: Quantitative evaluation and qualitative inferences, *Genome Res.*, 10:1024–1210, 2000.

[10] Liberles, D., Thoren, A., Heijne, G., and Elofsson, A., The use of phylogenetic profiles of gene predictions, *Current Genomics*, 3:131–137, 2002.

[11] Moreno-Hagelsieb, G. and Collado-Vides, J., A powerful non-homology method for the prediction of operons in prokaryotes, *Bioinformatics*, 18:S329–S336, 2002.

[12] Marcotte, E.M., Computational genetics: Finding protein function by nonhomology methods, *Curr. Opin. Struct. Biol.*, 10:359–365, 2000.

[13] Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O., and Eisenberg, D., A combined algorithm for genome-wide prediction of protein function, *Nature*, 402:83–86, 1999.

[14] Palenik, B., Brahamsha, B., Larimer, F.W., Land, M., Hauser, L., Chain, P., Lamerdin, J., Regala, W., Allen, E.E., McCarren, J., *et al.*, The genome of a motile marine *Synechococcus*, *Nature*, 424:1037–1042, 2003.

[15] Prestridge, D., SIGNAL SCAN: A computer program that scans DNA sequences for eukaryotic transcriptional elements, *CABIOS*, 7:203–206, 1991.

[16] Pellegrini, M., Marcotte, E., Thomopson, M., Eisenberg, D., and Yeates, T., Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles, *Proc. Natl. Acad. Sci.*, 96:4285–4288, 1999.

[17] Pavlidis, P., Weston, J., Cai, J., and Noble, W.S., Learning gene functional classifications from multiple data types, *J. Comput. Biol.*, 9(2):401–411, 2002.

[18] Riley, M., and Labedan, B., *E. coli* gene products: Physiological functions and common ancestries, *Escherichia coli and Salmonella: Cellular and Molecular Biology*, 2nd edition, In Neidhardt, F., *et al.*, editors, ASM Press, Washington, DC, 2118–2202, 1996.

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

[20] 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.*, 30:2886–2893, 2002.

[21] Salgado, H., Santos-Zavaleta, A., Gama-Castro, S., Millan-Zarate, D., Blattner, F.R., and Collado-Vides, J., RegulonDB (version 3.2): Transcriptional regulation and operon organization in *Escherichia coli* K-12, *Nucleic Acids Research*, 29(1):72–74, 2001.

[22] Tatusov, R., Koonin, E., and Lipman, D., A genomic perspective on protein families, *Science*, 278:631–637, 1997.

[23] Vert, J., A tree kernel to analyze phylogenetic profiles, *Bioinformatics*, 18:s276–s284, 2002.

[24] Thompson, D., Xu, Y., and Tidge, J., *Microbial Functional Genomics*, Wiley-LISS, 2004.

[25] 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.

[26] Zheng, Y., Szustakowski, J.d., Fortnow, L., Roberts, R.J., and Kasif, S., Computational identification of operons in microbial genopmes, *Genome Res.*, 12:1221–1230, 2002.