

Phylogenetics

Combining phylogenetic motif discovery and motif clustering to predict co-regulated genes

Shane T. Jensen^{1,*}, Lei Shen^{2,†} and Jun S. Liu²¹Department of Statistics, The Wharton School, University of Pennsylvania, PA, USA and ²Department of Statistics, Harvard University, Cambridge, MA, USAReceived on April 29, 2005; revised on July 22, 2005; accepted on August 11, 2005
Advance Access publication August 16, 2005**ABSTRACT**

Motivation: We present a sequence-based framework and algorithm PHYLOCLUS for predicting co-regulated genes. In our approach, *de novo* discovery methods are used to find motifs conserved by evolution and then a Bayesian hierarchical clustering model is used to cluster these motifs, thereby grouping together genes that are putatively co-regulated. Our clustering procedure allows both the number of clusters and the motif width within each cluster to be unknown.

Results: We use our framework to predict co-regulated genes in the bacterium *Bacillus subtilis* using six other closely related bacterial species. Our predicted motifs and gene clusters are validated using several external sources and significant clusters are examined in detail. An extension to the discovery and clustering of two-block motifs can be used for inference about synergistic binding relationships between transcription factors.

Availability: Software and Supplementary Materials can be downloaded at <http://stat.wharton.upenn.edu/~stjensen/research/phyloclus.html> or <http://www.fas.harvard.edu/~junliu/phyloclus.html>

Contact: stjensen@wharton.upenn.edu

1 INTRODUCTION

Genes are often regulated in living cells by proteins called transcription factors (TFs) that bind directly to short segments of DNA in close proximity to their target genes. These short segments have a conserved appearance, which we call a motif, shared by each binding site of the TF. The determination of TF binding sites (TFBSs) can be done experimentally through a labor-intensive process known as ‘footprinting’, but an attractive alternative is to use a combination of certain genomic and computational approaches. For example, microarray experiments can be done at various conditions and genes with similar expression profiles can be clustered. Then a statistically based algorithm such as BioProspector (Liu *et al.*, 2001) or AlignACE (Roth *et al.*, 1998) can be applied to search the upstream regions of a set of co-expressed genes for enriched motif signals. For a review of statistically based *de novo* motif finding algorithms see Jensen *et al.* (2004). However, microarray data collection is expensive and many co-expressed gene clusters are quite heterogeneous in terms of their regulation mechanisms.

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

When the sequence information from several closely related species is available, an alternative motif discovery strategy is to look for binding sites that are conserved among sets of orthologous genes across different species, rather than across different genes within the same species. This ‘phylogenetic footprinting’ strategy operates under the assumption that biologically important DNA sequence features such as TFBSs are likely to be conserved by evolution. Phylogenetic footprinting has the advantage that clusters of co-regulated genes do not have to be inferred beforehand (e.g. by microarray data), since we are looking for motifs that are conserved for a particular gene across species instead of across genes within a single species. Any motif that is unique to a particular species will not be detected by this method. Another restriction of this method is that the complete sequence information from several related species must be known and orthologous genes within these species must be identified. Fortunately, the genomes of many species have been completely sequenced and are available publicly (e.g. NCBI, www.ncbi.nlm.nih.gov). McCue *et al.* (2001) used the sequence information from nine bacterial species to identify TF-binding sites in *Escherichia coli*. We apply our procedure to the bacterium *Bacillus subtilis*, which also has several related species for which complete genome information is available.

Building on top of the concept of phylogenetic footprinting is the idea that many of the motifs discovered within each of these orthologous upstream regulatory sequences will be similar enough in appearance that we will be able to group them into clusters. If the motifs found upstream of several *B.subtilis* genes are similar enough to be clustered together, then it is possible that the same TF (recognizing that common motif) is targeting each of the genes in that cluster. Thus, by combining statistical techniques for both motif discovery and motif clustering, one can infer potentially co-regulated gene clusters (Qin *et al.*, 2003). However, the earlier motif clustering method of Qin *et al.* (2003) was separate from the phylogenetic motif discovery process, making it less attractive to both practitioners and theoreticians.

Here we present a systematic framework and its associated algorithm that combines both motif discovery and motif clustering. As described in Sections 2.1 and 2.2, our PHYLOCLUS framework builds upon the techniques presented by McCue *et al.* (2001) and Qin *et al.* (2003), but with several novel generalizations that result in a more principled procedure. Our motif discovery techniques are less restrictive in several ways, most notably allowing for a variable motif width and unknown motif abundance. Our clustering procedure is more flexible than traditional clustering schemes (Hartigan,

1975) since it avoids the use of arbitrary thresholds and allows for an unknown number of clusters. Another technical advance of our procedure is the flexibility to allow unknown motif widths that can vary between clusters. In addition to the discovery and clustering of single-block motifs, our procedure is also extended in Section 2.3 to the discovery of two-block motifs with a variable-length gap, thereby allowing for dimer motifs and the synergistic binding of two TFs in close proximity to each other. In Section 2.4, we present several validation strategies for evaluating the performance of PHYLOCLUS when applied to the bacterium *B.subtilis*. The results from our *B.subtilis* application are presented and validated in Section 3, along with detailed examination of several clusters. Finally, we discuss our results in Section 4 and suggest several areas for further development.

2 MATERIALS AND METHODS

Our procedure begins with a set of N genes in a particular species of interest, and the corresponding set of orthologous genes in several related species. The responsibility for identifying these orthologous genes is left to the user, although our procedure for *B.subtilis* is briefly discussed in Section 3.1. The user is also responsible for organizing each gene of interest and its orthologues into N orthologous gene sets (OGSs) and collecting the upstream regulatory sequence for each gene in the OGS.

2.1 Phylogenetic motif discovery

The upstream regulatory regions for each OGS form a small sequence dataset, which we hypothesize contains multiple different TF-binding motifs that have been conserved by evolution. The motif discovery component of PHYLOCLUS involves the repeated use of the programs, BioProspector (Liu *et al.*, 2001) and BioOptimizer (Jensen and Liu, 2004), to find multiple unique motifs in the same sequence dataset. BioOptimizer uses a scoring function to compare and optimize motifs discovered by any motif-finding program, such as BioProspector.

For a particular OGS sequence dataset, the following motif discovery procedure is applied to find conserved one-block motifs. All parameters of this procedure can be specified by the user of PHYLOCLUS (as an example, we indicate the parameters used in our *B.subtilis* application).

- (1) The motif-finding program BioProspector (Liu *et al.*, 2001) is used to find one-block motifs of a given width. Since the motif width must be pre-specified for BioProspector, the program is run separately for n user-specified suggestions for different motif widths (e.g. 8, 10, . . . , 30 bp). For each width, the top k motifs are collected, with k also being specified by the user (e.g. $k = 5$).
- (2) Since BioProspector is a stochastic algorithm, independent runs of the program may give different results. To account for this fact, Step 1 can be repeated m times (e.g. $m = 3$) for each suggested width, resulting in a total of $n \times m \times k$ BioProspector motifs, many of which might be identical or very similar. Redundant motifs (motifs with completely identical predicted sites) are removed by PHYLOCLUS.
- (3) Each non-redundant motif is separately scored and optimized using the program BioOptimizer (Jensen and Liu, 2004), which also allows the motif width to vary in order to find both the optimal motif width and optimal predicted binding sites. The motif with the highest BioOptimizer score is retained as the ‘best motif’.
- (4) BioOptimizer also calculates a ‘null score’ based entirely on the background sequence of the dataset that serves as a simple diagnostic measure for a discovered motif. If the best motif has a lower score than the null score, it was removed from consideration. Otherwise, the motif was retained for the motif clustering component of PHYLOCLUS.

The retained motif is then ‘masked out’ of the sequence dataset by replacing all its binding sites with N.

- (5) With this new ‘masked’ sequence dataset, the entire motif-finding procedure (Steps 1–4) is repeated for a user-specified number of times or until no more motifs are found that have a BioOptimizer score greater than the null score. A similar iterative-masking approach is also implemented by Roth *et al.* (1998) in their program AlignACE.

Applying this iterative-masking one-block motif discovery strategy to each OGS sequence dataset separately results in several discovered one-block motifs (summarized as count matrices) associated with each orthologous gene set. Figure 1 provides a summary of the motif discovery component of PHYLOCLUS for our *B.subtilis* application.

2.2 Bayesian clustering of discovered motifs

There are several traditional statistical techniques for clustering observations. Hierarchical Tree Clustering joins observations together into successively larger clusters based upon some sort of similarity measure between observations that is specified by the user. The use of a similarity measure assumes that the observations are fixed and known, which is not true in this case of our estimated motifs. In addition, the result of this algorithm is a tree that joins all observations together, and it is not clear where the tree should be ‘cut’ in order to produce a set of clusters. K -means clustering groups observations into a pre-determined number of clusters by minimizing some sort of within-cluster distance measure. However, the number of clusters can be difficult to be determined as we have very little idea how many motif clusters we should expect.

We use a Bayesian hierarchical model to infer the clusters within our collection of motifs, which are represented by count matrices Y_{ijk} , where i indexes the motif, j indexes the column within each motif and k indexes the four possible nucleotides within each column. Our within-motif level model is: $p(\mathbf{Y}_i | \Theta_i) = \prod_{j=1}^w p(\mathbf{Y}_{ij} | \theta_{ij})$, where

$$\mathbf{Y}_{ij} = (Y_{ija}, \dots, Y_{ijt}) \sim \text{Multinomial}(N_i, \theta_{ij} = (\theta_{ija}, \dots, \theta_{ijt})),$$

whereas our between-motif level model is

$$\Theta_i = (\theta_{i1}, \dots, \theta_{iw}) \sim F(\cdot),$$

where $F(\cdot)$ is an unknown distribution that is assumed to follow a Dirichlet process. This model enables similar motifs to be clustered together into groups with identical frequency matrices (see Supplementary Materials for details). The hierarchical structure lets us account for uncertainty in the alignment matrix that represents each TF motif (by assuming a product multinomial distribution), whereas most clustering programs would require the motif matrix to be known without error.

We implement our model via a Gibbs sampling algorithm, which iteratively samples unknown parameters (or sets of parameters) one at a time by conditioning on the current values of all the other parameters. Using the notation, $z_i = c$ if the i -th motif is in the c -th cluster and $z_i = 0$ if the i -th motif is unclustered, then each iteration of our algorithm consists of choosing a destination for each motif i based on the current set of clusters \mathbf{z}_i (containing all motifs except the i -th one). This choice is between leaving the i -th motif unclustered, with probability

$$p(z_i = 0 | z_{-i}, \mathbf{Y}) \propto \frac{1}{n} \prod_{j=1}^w \frac{\prod_k \Gamma(Y_{ijk} + \alpha)}{\Gamma(\sum_k Y_{ijk} + 4\alpha)} \frac{\Gamma(4\alpha)}{\Gamma(\alpha)^4}$$

or placing the i -th motif in an existing cluster c (with total count matrix of \tilde{Y}_c), with probability

$$p(z_i = c | z_{-i}, \mathbf{Y}) \propto \frac{n_c}{n} \prod_{j=1}^w \frac{\prod_k \Gamma(Y_{ijk} + \tilde{Y}_{cjk} + \alpha)}{\Gamma(\sum_k Y_{ijk} + \tilde{Y}_{cjk} + 4\alpha)} \frac{\Gamma(\sum_k \tilde{Y}_{cjk} + 4\alpha)}{\prod_k \Gamma(\tilde{Y}_{cjk} + \alpha)},$$

where n_c is the current size of the cluster c and α represents prior pseudo-counts that are added to each count matrix. A complete iteration of our Gibbs sampling algorithm results in a complete sample \mathbf{z} of our cluster indicators,

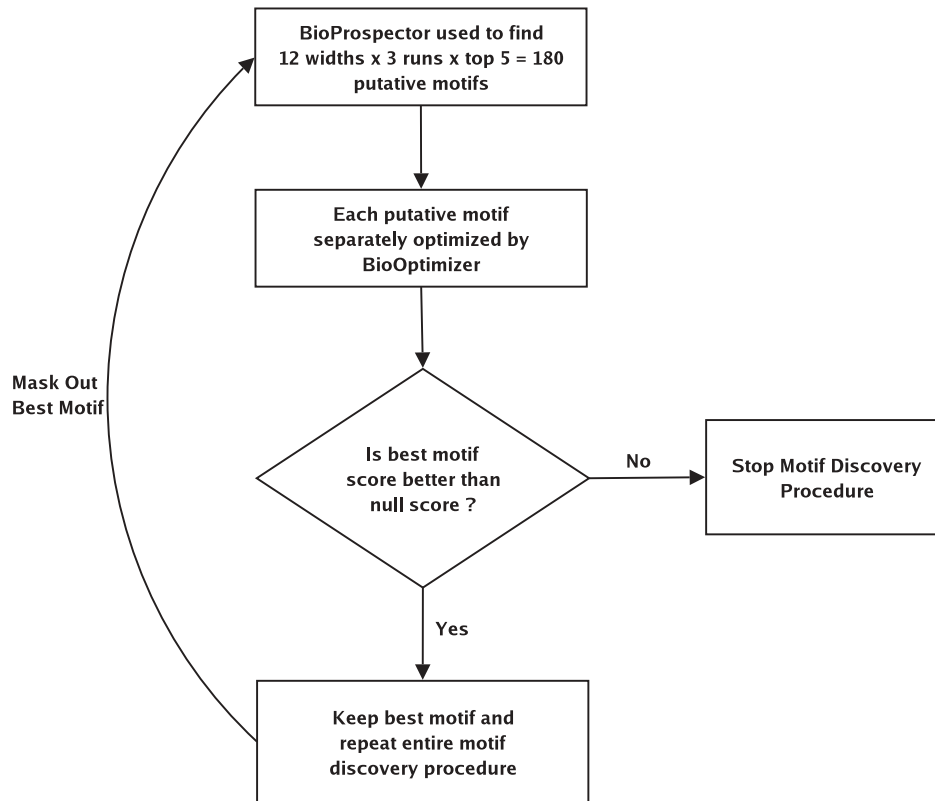


Fig. 1. Flowchart for motif discovery procedure.

which also represents a complete partition of our motif matrices. A key benefit of this implementation is that it allows not only the clusters themselves to vary (in terms of which motifs are clustered together) but also the number of clusters to vary from iteration to iteration.

An additional advantage of PHYLOCLUS is a user-specified option to run a ‘variable-width’ clustering procedure in which the motif width within each cluster is allowed to vary, so that we have widths w_c ($c = 1, \dots, C$, where C is the current number of clusters) instead of a single common width w . This novel component is beneficial because the width of TFBSs is often unknown and can be quite different for different TFs. Using a Poisson (λ) prior distribution to model each w_c , we can add an additional step to our Gibbs sampling algorithm that, for each cluster c , chooses a new width w_c with probability

$$p(w_c | z_c) \propto \prod_{j=1}^{w_c} \frac{\Gamma(\tilde{Y}_{cjk} + \alpha)}{\Gamma(\sum_k \tilde{Y}_{cjk} + 4\alpha)} \frac{\Gamma(4\alpha)}{\Gamma(\tilde{Y}_{cjk})^4} \cdot \prod_k \theta_{0k}^{B_{ck}} \frac{\lambda^{w_c} e^{-\lambda}}{\Gamma(w_c)},$$

where \tilde{Y}_{cjk} and B_{ck} are, respectively, the motif and background nucleotide counts in the cluster c , and θ_0 are the background probabilities for each nucleotide. Alternatively, the user can run a ‘fixed-width’ version of the clustering procedure, in which case the common motif width w is specified by the user a priori.

The resulting ‘best clusters’ from either our fixed-width or variable-width clustering procedures are the set of clusters \hat{z} that give the highest overall posterior probability, which for the fixed-width model is

$$p(\hat{z} | \mathbf{Y}) \propto \prod_{l=1}^L \prod_{j=1}^w \frac{\Gamma(\tilde{Y}_{jk} + \alpha)}{(\sum_k \tilde{Y}_{jk} + 4\alpha)} \times \frac{\prod_{l=1}^L (n_l - 1)!}{n!}.$$

The posterior value of \mathbf{z} for our variable-width model is similar, but has additional terms for the background nucleotide counts \mathbf{B} and variable widths w_c .

We measure the strength of each cluster by calculating the logarithm of the Bayes factor (Kass and Raftery, 1995) for the current cluster (details given in the Supplementary Material). The clusters within our best partition can then be ranked by this measure of cluster strength, giving us an extra measure of confidence/uncertainty about inferred clusters. We can also measure clustering strength by calculating, for each motif, the posterior probability that it should belong to that cluster, as opposed to any of the other clusters. The user can utilize these individual clustering probabilities to filter the best partition clusters and remove any motifs that are only weakly associated with their respective clusters.

2.3 Extension to two-block motifs

Since many transcription binding sites are composed of two ungapped blocks, with a variable-length gap in between, our PHYLOCLUS algorithm also contains an option for the discovery and clustering of two-block motifs. This two-block specification may also detect binding sites from two separate TFs that bind in close proximity to each other. PHYLOCLUS uses BioProspector and a strategy similar to the one described in Section 2.1 to find two-block motifs, except that the user must also specify a suggested range of gap widths (e.g. 12–15 bp) in addition to suggested widths of the motifs in each of the two blocks (e.g. 8–8, 10–10, ..., 20–20). Each discovered motif is then optimized by a two-block version of BioOptimizer, which finds the optimal motif width and set of predicted sites. We examine two strategies for clustering the discovered two-block motifs in our *B.subtilis* application. First, an independent-block strategy separates a two-block motif into two independent single block motifs, and clusters these new single block motifs together with the original one-block motifs (m one-block motifs + n two-block motifs $\rightarrow m + 2n$ independent-block motifs). This strategy ignores the linkage between the two blocks, but allows both the two-block and one-block motifs to be clustered together. Alternatively, a joint-block strategy clusters the two-block motif as a single entity, which acknowledges the inherent link

between the two blocks, but does not allow the two-block motifs to be clustered together with the one-block motifs.

2.4 Validation of predicted gene clusters

To evaluate our predicted co-regulated gene clusters, we constructed four validation measures based upon external information. Functional Category Over-Representation examines whether or not the predicted clusters contain genes with the same function. Each *B.subtilis* gene has been classified into a set of functional categories, which are available on the Subtilist website (Moszer *et al.*, 1995). GeneMerge (Castillo-Davis and Hartl, 2003) was used to calculate a *P*-value (from a Hypergeometric distribution with a Bonferroni correction) for the over-representation of each functional category in a given cluster. Known TF Over-Representation examines whether the predicted clusters contain genes that are known to be controlled by the same TF protein based on a list of 650 known TF-gene interactions from the DBTBS database (Makita *et al.*, 2004), although this list presumably catalogues only a miniscule fraction of the true gene-TF interactions. Again, GeneMerge was used to calculate a *P*-value for the over-representation of interactions with a particular TF in a given cluster.

We also use gene expression patterns within predicted clusters to see if genes within particular clusters are co-expressed. Our expression dataset consists of ratios of differential expression on cDNA microarrays from eight different experimental conditions in *B.subtilis* [Conlon *et al.* (2004) and Eichenberger, Wang and Losick, unpublished data]. Two different measures of microarray co-expression were considered: Median Within-Cluster Correlation *S* and Within-Cluster Variance Ratio *T*. The Pearson correlation was calculated between each possible set of two genes in a particular cluster, and the median value of these correlations is our measure *S*. The absolute value of the correlation was actually used in *S* to allow for genes in the same cluster that are regulated by the same TF but in opposite ways (one repressed while the other is enhanced). The measure *T* was calculated as the ratio of the within-cluster variance to the total-cluster variance among all genes in the dataset. For a cluster with *k* out of *n* total genes, this ratio is

$$T = \frac{\frac{1}{8} \sum_{i=1}^8 \left[\frac{1}{k} \sum_{j=1}^k (x_{ij} - \bar{x}_i^*)^2 \right]}{\frac{1}{8} \sum_{i=1}^8 \left[\frac{1}{n} \sum_{j=1}^n (x_{ij} - \bar{x}_i)^2 \right]},$$

where x_{ij} is the differential expression ratio for the gene *j* in the experimental condition *i*, x_i and \bar{x}_i^* are the differential expression ratios in the experimental condition *i* averaged over all genes in the dataset and only the genes in the cluster, respectively. If PHYLOCLUS has been effective, we expect low values of *T* and high values of *S*. We estimate *P*-values for our observed *S* and *T* by comparing them to values of *S* and *T* from randomly generated clusters of the same size. The main weaknesses of these expression-based measures are the limited number of experimental conditions present in our dataset, as well as the inherent noise present in the microarray data. Despite the fact that each validation measure has particular limitations, the use of several measures simultaneously should give us a good indication of the effectiveness of PHYLOCLUS when applied to *B.subtilis*.

3 RESULTS

3.1 Collection of orthologous gene sets

The annotated genome sequences for *B.subtilis* and six other bacterial species were acquired from the NCBI website (<http://www.ncbi.nlm.nih.gov>). To avoid genes in the middle of operons, only genes with an upstream intergenic region of more than 50 bp were considered to be ‘valid’. For each of these valid *B.subtilis* genes, orthologous genes in the other species were identified using a reciprocal BLAST best-hit procedure (Remm *et al.*, 2001), with a significance threshold of 10^{-10} . Table 1 gives the number of total and valid genes for each species, along with the number of valid and orthologous genes between each species and *B.subtilis*.

Table 1. Bacterial species included in the study

Species and reference	Genome size (Mb)	Total genes	Valid genes	Valid and orthologous genes
<i>Bacillus anthracis Ames</i>	5.2	5311	3717	1267
<i>Bacillus halodurans</i>	4.2	4066	2634	1096
<i>Bacillus subtilis</i>	4.2	4225	2716	—
<i>Clostridium acetobutylicum</i>	3.9	3672	2303	582
<i>Clostridium perfringens</i>	3.0	2660	1863	515
<i>Listeria innocua</i>	3.0	2968	1659	702
<i>Oceanobacillus ihenysis</i>	3.6	3500	2259	1029

For each gene in *B.subtilis* with at least one orthologue, the *B.subtilis* gene and its orthologues were collected into an OGS, giving us a total of 1516 OGSs. We collected the regulatory sequence for each gene in each OGS, which was defined as the sequence (up to a maximum of 500 bp) upstream of the translation start site. This maximum length was chosen so that we could safely assume that our sequences contained the entire upstream regulatory region of each gene in our OGSs, since the length of upstream regulatory regions tend to only be a couple of hundred basepairs in bacteria. For applications in more complex organisms, the user may want to include a larger length of upstream sequence. Our sequences were also restricted in each case to be intergenic (no overlap with any coding regions). A second dataset considered was a subset of 172 OGSs for which reliable *B.subtilis* TF binding site information was available from the DBTBS online database Makita *et al.* (2004). This ‘studysset’ was used for additional validation of the motif discovery component of PHYLOCLUS.

3.2 Discovery of phylogenetically conserved motifs

The motif discovery component of PHYLOCLUS was applied to our 1516 OGS sequences datasets, resulting in 771 one-block predicted motifs. In addition, we applied the optional two-block motif discovery component of PHYLOCLUS, resulting in 1443 two-block predicted motifs. For validation, we focussed on the predicted binding sites for the subset of 95 one-block motifs and 170 two-block predicted motifs from our 172 studysset OGSs, for which we have additional information. Among the *B.subtilis* genes in these OGSs, we have 241 confirmed binding sites belonging to 44 TFs from the DBTBS database Makita *et al.* (2004), for which we divided the sites into two categories: a ‘conserved’ subset consisting of sites that are also present in some of orthologous upstream sequences of the other species, and the remaining ‘non-conserved’ sites, which were not present in the orthologous upstream sequences of the other species. Details of this categorization are given in the Supplementary Material.

Table 2 gives the specificity and sensitivity for our discovered binding sites relative to the 241 confirmed binding sites from the DBTBS database, which were also subdivided into one-block and two-block categories. Our one-block binding sites had a similar specificity compared with the two-block sites, but the two-block binding sites had a higher sensitivity. The specificity and sensitivity of our discovered binding sites were much higher for the ‘conserved’ subset of known motif sites. This result was expected, since our motif discovery method relies on the assumption that

Table 2. Prediction specificity and sensitivity for our motif discovery procedure

Set of known sites	Number of matches	Number of predicted sites	Specificity	Number of known sites	Sensitivity
One-block motifs					
All	58	220	0.2636	241	0.2407
Conserved	45	112	0.4018	74	0.6081
Non-conserved	13	112	0.1161	167	0.0778
Two-block motifs					
All	142	548	0.2591	241	0.5892
Conserved	65	177	0.3672	74	0.8784
Non-conserved	77	403	0.1911	167	0.4611

TF-binding sites are conserved across different species, which is satisfied by our ‘conserved’ subset of known motif sites, but not the ‘non-conserved’ subset.

Our discovered binding sites that did not match the experimental sites could be either false positives or true sites that have not yet been experimentally characterized.

In our Supplementary Materials, we present similar validation results for a set of experimentally confirmed binding sites from the σ^E , σ^F , σ^H and σ^H regulons in *B.subtilis* (Eichenberger *et al.*, 2003, 2004).

3.3 Clustering of discovered motifs

We present several different sets of clustering results (summarized in Table 3) based on the various optional analyses that are possible using PHYLOCLUS. Both independent-block and joint-block clustering strategies were implemented. A fixed motif width of 12 bp was assumed for the one-block case and 8 bp for each block in the two-block case. In addition, we examine clustering results on our collection of one-block motifs under the optional extension of PHYLOCLUS that allows the motif width within each cluster to be unknown and vary between clusters. For each dataset and clustering strategy, our analysis focused on the set of best clusters, which we defined as the best partition of clusters generated by PHYLOCLUS that have been ‘filtered’ to remove any motifs with individual clustering probabilities of <0.75 . Details and the full list of best clusters for each dataset and clustering strategy are given in the supplementary material. Table 3 gives, for each dataset and clustering strategy, the proportion of best clusters that were statistically significant (at level $\alpha = 0.05$) for each validation measure (see Section 2.4), as well as the proportion of best clusters that were significant across multiple validation measures.

At a significance level of 0.05 and under a null hypothesis that our clusters are not biologically relevant, we would expect $\sim 20\%$ significance on at least one significance measure and $<3\%$ significance on multiple measures (with the additional assumption that each validation measure is independent). For each dataset and clustering strategy, we observed a much higher rate of significant clusters on at least one measure and on multiple measures than would be expected by chance, indicating that each version of PHYLOCLUS has produced some clusters that are biologically relevant. It is not surprising to see that the proportion of significant genes is lower for our genome clusters than our studyset clusters, since the criterion for

including a gene in our studyset was that the gene must have a known TF interaction, and known TF interactions is also one of our validation measures. In fact, when we break down our percentage of significant clusters into the individual validation measures, we see that the largest difference between the studyset and whole genome results is definitely the known TF validation measure. It is also interesting to note that the fixed-width and variable-width versions of the one-block clustering are quite similar in terms of the percent significant on at least one validation measure, despite differences in terms of the number of clusters and the average cluster size. A general trend seems to be that the variable-width clustering produces larger numbers of clusters, but that these clusters are weaker, which is indicated by the fact that the number of clusters is dramatically reduced after filtering for weakly clustered motifs, resulting in a smaller number of variable-width clusters. The total number of motifs differ between fixed-width and variable-width clustering results because assuming a fixed width of 12 bp excludes many shorter discovered motifs.

3.4 Comparison with randomly permuted orthologous gene sets

In order to check our assumption that motifs conserved by evolution are likely to be functionally relevant, we implemented our one-block fixed-width version of PHYLOCLUS on our studyset dataset, but with the upstream sequences randomly permuted between different OGSs so that any existing phylogenetic relationship within each OGS is eliminated. Details of our randomization procedure are given in the Supplementary Materials. The results from these randomly permuted OGSs are clearly inferior to our original analysis in several ways. Far fewer conserved motifs (61 discovered motifs) were discovered in these random OGSs, compared with 298 discovered motifs from the original studyset. When checked against our 241 confirmed binding sites from the DBTBS database (Makita *et al.*, 2004), our discovered motifs from the randomized OGSs showed a much lower sensitivity (0.0871) than the sensitivity from our original analysis (0.2407). It is also interesting that the low sensitivity for the random OGSs is similar to the low sensitivity (0.0778) from our original analysis of only the ‘non-conserved’ cases of known binding sites. The discovered motifs from these random OGSs also clustered together into a much lower number of clusters (5 independent-block clusters) versus the original studyset (97 independent-block clusters) and none of these five clusters was significant on multiple validation measures.

3.5 Detailed examination of some significant clusters

All predicted clusters that were significant on multiple validation measures are shown in Table 4, ranked by cluster strength within each dataset and clustering strategy.

One interesting subset has the multiply-significant clusters S-Ind-Fix-4, S-Ind-Fix-5 and S-Ind-Var-1, all of which have significant over-representation of the functional category Nucleotide Metabolism and the known TF PurR. Combined together, these three clusters contain the genes *dra*, *purA*, *purR*, *purE*, *recA*, *ytiP* and *yumD*. Saxild *et al.* (2001) states that the PurR TF is involved in the purine biosynthetic pathway in *B.subtilis*, and that PurR binds to five genes (*purR*, *purE*, *ytiP*, *yumD* and *purA*), all of which are contained in our predicted clusters. Saxild *et al.* (2001) also suggest that the PurR TF recognizes a two-block motif with a CGAA first block and a TTCC second block, which also match the highly conserved portions of the

Table 3. Clustering results from PHYLOCLUS for two different OGs (studysset versus genome), and three different clustering strategies (fixed-width independent-block, variable-width independent-block and fixed-width joint-block)

Statistic	Studysset OGSS		Joint fixed-width	Genome OGSS		Joint fixed-width
	Independent Fixed-width	Variable-width		Independent Fixed-width	Variable-width	
Total number of motifs	298	433	153	2543	3657	1278
Number of clusters	97	115	34	719	1111	300
Number of filtered clusters	64	42	30	568	138	200
Average cluster size	2.38	3.40	2.47	2.96	2.81	2.22
Percentage of significant clusters						
Functional categories	6.25	9.52	13.33	12.68	10.14	6.50
Known TF interactions	21.88	14.29	23.33	2.46	1.45	2.50
Expression median correlation	3.13	7.14	10.00	6.16	3.62	4.00
Expression variance ratio	6.25	4.76	13.33	4.40	6.52	5.50
Percentage of significant clusters						
At least one measure	29.69	30.95	40.00	21.30	19.57	15.50
Multiple measures	7.8	4.76	10.00	3.70	2.17	3.00

consensus sequences from our predicted clusters S-Ind-Fix-4 (a CGAA motif), S-Ind-Fix-5 (a TTCG motif) and S-Ind-Var-1 (a TTCG motif). The cluster S-Jnt-Fix-1 is not significant on multiple measures but also contains four of these genes and has a matching consensus sequence CGAAcatT--AatgTTCG, which combines the motif signals of the one-block clusters S-Ind-Fix-4 and S-Ind-Fix-5. Although they also are not significant on multiple measures, several whole-genome clusters (G-Ind-Fix-36, G-Ind-Fix-61, G-Ind-Fix-153 and G-Jnt-Fix-1) share many of the same PurR-controlled genes also found within the studysset clusters. It is worth noting that these PurR clusters do not show similarity of gene expression on either the variance or the correlation measure introduced in Section 2.4, which demonstrates that our procedure can cluster groups of co-regulated genes that would not be detected by standard procedures based entirely on gene expression data.

In many other cases, similar clusters were predicted by both the independent and joint-block procedures, but some genes are only found in either the independent or the joint-block clusters. This might indicate that the additional independent block genes are bound by a TF that only resembles a portion of the joint-block motif. Another explanation for this behavior would be that the joint-block motif in some of these cases is not a true two-block binding motif, but rather consists of binding sites for two single-block motifs that occur in close proximity to one another in each of the genes in the joint-block cluster. In this case, the additional independent-block motifs would represent genes that are bound by only one of those TFs, but not the other, and so only are included in an independent-block cluster but not the joint-block cluster. One case we examined was the specific clusters S-Ind-Fix-9 (ycdH, yciC) and S-Jnt-Fix-4 (ycdH, yciC, dhbA). Gaballa *et al.* (2002) analyzed the Zur regulon and demonstrated that the genes yciC and ycdH are bound by the Zur TF. They describe Zur as a regulator of zinc uptake, which confirms the over-representation of Transport/Binding proteins in the S-Ind-Fix-9 cluster. Gaballa *et al.* (2002) also presented a 28 bp long consensus sequence for the Zur-binding motif AAttTAAATCGTAATcATTacGaTTTAA based on four genes and noted that the central region of this consensus sequence TAATnATTA is shared by two other TFs, PerR and Fur. We see this same consensus sequence AATcATTA in our S-Ind-Fix-9 cluster,

which seems to support the theory that the additional joint block gene (*dhbA*) has a binding motif resembling the central region of the Zur motif, but does not have the entire Zur motif. According to the DBTBS database (Makita *et al.*, 2004), *dhbA* is known to be bound by Fur, which further confirms this hypothesis. A second case is the genome clusters G-Ind-Fix-6 and G-Jnt-Fix-3 that share the genes *ylxY* and *yunB*, along with several non-common genes in both the clusters. The cluster G-Jnt-Fix-3 is significant on the functional category over-representation (Sporulation) and over-representation of a known TF, σ^E . The consensus sequences of G-Ind-Fix-6 (ttgaAGgAggg) and G-Jnt-Fix-3 (cCccctCt--AggggggG) have portions that loosely match the known motif for the ribosomal-binding site, also known as the Shine-Dalgarno sequence (Shine and Dalgarno, 1974), which is known to bind in close proximity to σ TFs. One explanation could be that the joint block motif G-Jnt-Fix-3 is actually a combination of two motifs: portions of a σ TF-binding motif combined with the ribosomal-binding site.

4 DISCUSSION

Our PHYLOCLUS framework that combines phylogenetic motif discovery with principled motif clustering should prove useful for predicting co-regulated genes in many organisms. Our motif discovery component involved the combination of a stochastic Gibbs sampling-based motif-finding program, BioProspector (Liu *et al.*, 2001), and a deterministic optimization algorithm, BioOptimizer (Jensen and Liu, 2004), both of which are based on a Bayesian motif model (Jensen *et al.*, 2004). BioOptimizer not only optimizes the signal for fixed-width motifs, but also allows the motif width to be optimized, which is an improvement over previous phylogenetic footprinting methods, such as Qin *et al.* (2003), which used flanking sequences around discovered fixed-width motifs to improve performance. The clustering component of PHYLOCLUS allows the number of clusters to be unknown and includes an optional extension that allows the motif width within each cluster to vary. PHYLOCLUS can also be used to discover two-block motifs with a variable gap, which is the form taken by many bacterial TF-binding motifs. In addition, this two-block option allows for

Table 4. Statistics for clusters that are significant on multiple validation measures

Cluster	Width	Strength	Size	Consensus	P-values				Significant Function	TF
					EC	EV	FC	TF		
S-Ind-Fix 4	12	82.6	3	AaaaCGAACAtT			0.006	0.000	Metabolism-nucs	PurR
S-Ind-Fix 5	12	77.8	3	AAtgTTCGtaTT			0.011	0.000	Metabolism-nucs	PurR
S-Ind-Fix 6	12	76.3	3	GaAAgCGcTTtC		0.046		0.006		CcpA
S-Ind-Fix 9	12	60.4	2	CGTAATcATTAC		0.009	0.000		Transport/bindi	Zur
S-Ind-Fix 39	12	28.5	2	AaAagtAtATGt	0.037			0.030		SigE
S-Ind-Var 1	6	120.3	6	TgTTCCG			0.003	0.000	Metabolism-nucs	PurR
S-Ind-Var 40	6	13.0	2	tATttg	0.008			0.030		SigE
S-Jnt-Fix 7	8-8	78.4	3	cccctcCt-GgaggagA			0.011	0.005	Sporulation	SigE
S-Jnt-Fix 18	8-8	39.6	2	AtatTTTT-aaAGgata	0.034			0.030		SigE
S-Jnt-Fix 19	8-8	37.2	2	GgcAacTc-TtcaAgTC		0.018		0.015		AbrB
G-Ind-Fix 25	12	168.1	4	AACatAtGTTCg		0.003		0.003		ComK
G-Ind-Fix 35	12	146.9	4	aAtaTtaCTTGa	0.024		0.017		Protein-synthes	
G-Ind-Fix 140	12	75.3	3	AgaCgaaTGtCT			0.027	0.006	Metabolism-carb	CcpA
G-Ind-Fix 145	12	70.6	3	aAgtgGaaagga			0.027	0.003	Metabolism-carb	CcpA
G-Ind-Fix 174	12	63.1	3	aTttagAcaAAA	0.011	0.003				
G-Ind-Fix 206	12	59.8	3	aaAggagagGAg	0.006	0.002				
G-Ind-Fix 258	12	55.6	3	caatttTcgACA	0.030		0.038		RNA-synthesis	
G-Ind-Fix 268	12	54.3	2	TGTCAaGACAtc		0.001	0.002		Metabolism-coen	SigA
G-Ind-Fix 281	12	53.4	2	CTTGaCatcaaT		0.001	0.002		Metabolism-coen	SigA
G-Ind-Fix 282	12	53.4	2	AtaaatGtCAAG		0.001	0.002		Metabolism-coen	SigA
G-Ind-Fix 283	12	53.3	3	aaatatAtatGT	0.037		0.022		Sporulation	
G-Ind-Fix 291	12	52.2	2	TttTttCACAt		0.001	0.000		Membrane-bioene	ResD
G-Ind-Fix 294	12	51.7	2	TAttaTaAtAaT		0.005	0.007		Metabolism-carb	SigA
G-Ind-Fix 300	12	49.8	3	taCaaagCAaat	0.001		0.022	0.004	Sporulation	SigE
G-Ind-Fix 304	12	47.4	2	tacgttataTtT		0.001	0.000		Membrane-bioene	ResD
G-Ind-Fix 345	12	33.7	2	ctATTtTagCAa	0.025	0.011	0.018		SimilartoBsub	
G-Ind-Fix 372	12	31.6	2	atCGtAgTaCgA	0.024	0.027				
G-Ind-Fix 397	12	30.6	2	TtgTacAAatga	0.021		0.014		Similartoother	
G-Ind-Fix 402	12	30.5	2	gcggtcGTggcg	0.042	0.022				
G-Ind-Fix 421	12	29.8	2	tGcggttagacA	0.007		0.000		Adaptation	
G-Ind-Fix 430	12	29.5	2	gaCAaatGccta		0.004	0.001		Sporulation	SigE
G-Ind-Fix 481	12	27.7	2	TCggtgAcTtcG		0.030	0.001		Protein-synthes	
G-Ind-Fix 548	12	24.5	2	ctTttaaGaaag	0.035		0.000			SigH
G-Ind-Var 26	6	62.0	4	GaAAaA	0.001	0.013				
G-Ind-Var 30	6	59.1	4	AccCTg	0.020		0.012		Cell-Wall	
G-Ind-Var 63	9	33.3	3	TttTAcCTC		0.036	0.000		DNA-replication	
G-Jnt-Fix 5	8-8	119.7	4	cCccctCt-AggggggG		0.002	0.000		Sporulation	SigE
G-Jnt-Fix 8	8-8	107.3	3	tCTTgACa-tGTcAAGa		0.003	0.006		Metabolism-coen	SigA
G-Jnt-Fix 34	8-8	66.6	2	TtttCACa-ttataTtT		0.001	0.000		Membrane-bioene	ResD
G-Jnt-Fix 37	8-8	46.4	2	TaTGTTcg-gctAtact	0.042	0.009				
G-Jnt-Fix 165	8-8	32.5	2	AAgtTaAt-GGAgAgAc	0.020	0.012				
G-Jnt-Fix 167	8-8	32.5	2	tcgtCAaa-cttgttgC	0.047	0.014				
G-Jnt-Fix 200	8-8	29.9	2	TcCTcCta-cAaGgAGg		0.022	0.001		Protein-synthes	

the potential discovery of motifs for pairs of TFs that bind in close proximity. Use of the Bayesian approach allows us to not only focus on a point estimate, or ‘best partition’ of clusters, but also account for variability within this best partition by evaluating the strength of each cluster and individual motif probabilities within each cluster. (Wang and Stormo, 2003) introduce an algorithm, Phylocon, which combines sequence information between related species with sequence information between co-regulated genes within a single species to improve motif discovery. Although it was not their intended goal, their framework (comparing motifs between genes that were discovered by cross-species sequence comparison) is somewhat similar to our strategy for inferring co-regulated

genes. In our application, we have used the whole-genome sequences of seven related bacterial species to discover TFBMs in the upstream regions of *B.subtilis* genes, and then have used similarities between these discovered motifs to predict possibly the co-regulated gene clusters. In addition to examining specific clusters for biological relevance, we were able to use external information to validate our entire set of clusters in a systematic fashion. Several clusters detected by our method (e.g. the purR clusters in Section 3.5) were significant but not similar on our gene expression measures, indicating that these clusters would not have been detected by methods based entirely on our gene expression data.

A potential improvement of our current method would be to incorporate the concept of evolutionary distances into our motif discovery procedures. Each sequence within a particular orthologous gene set was weighted equally with every other sequence by our motif-finding algorithms, despite the fact that these sequences came from different species with unequal phylogenetic distances between them. A more sophisticated motif discovery procedure should utilize this additional information to increase the power for detecting weaker motif signals. It is of interest to note that, for each of the six related species to *B.subtilis* that we used, the number of orthologous genes (Table 1) is <50% of the number of valid genes in any of the species. This level of phylogenetic distance is perhaps too far apart for the most efficient phylogenetic footprinting analysis, as indicated by McCue *et al.*, (2001, 2002). In their analysis, the closest and the second closest species to *E.Coli* have 92% and 68% orthologous genes, respectively. They confirmed that it is advantageous to include relatively closely related species, which is not always possible for a particular species of interest.

ACKNOWLEDGEMENTS

The authors thank the editor and the reviewers of this manuscript for their comments. The authors also thank Cristian Castillo-Davis for providing his *GeneMerge* program and for helpful discussions, Yuko Makita and Kenta Nakai for providing a flat file version of their DBTBS database, and Patrick Eichenberger and Richard Losick for helpful discussions about gene regulation in *B.subtilis*.

Conflict of Interest: none declared.

REFERENCES

Castillo-Davis,C. and Hartl,D. (2003) Genemerge—post-genomic analysis, data mining and hypothesis testing. *Bioinformatics*, **19**, 891–892.

- Conlon,E. *et al.* (2004) Determining and analyzing differentially expressed genes from CDNA microarray experiments with complementary designs. *J. Multivar. Anal.*, **90**, 1–18.
- Eichenberger,P. *et al.* (2003) The σ^E regulon and the identification of additional sporulation genes in *Bacillus subtilis*. *J. Mol. Biol.*, **327**, 945–972.
- Eichenberger,P. *et al.* (2004) The entire program of gene expression for a single differentiating cell type. *PLoS Biol.* accepted for publication.
- Gaballa,A. *et al.* (2002) Functional analysis of the *Bacillus subtilis* Zur regulon. *J. Bacteriol.*, **184**, 6508–6514.
- Hartigan,J. (1975) Clustering Algorithms. Wiley, NY.
- Jensen,S. and Liu,J. (2004) BioOptimizer: a Bayesian scoring function approach to motif discovery. *Bioinformatics*, **20**, 1557–1564.
- Jensen,S. *et al.* (2004) Computational discovery of gene regulatory binding motifs: a Bayesian perspective. *Stat. Sci.*, **19**, 188–204.
- Kass,R. and Raftery,A. (1995) Bayes factors. *J. Am. Stat. Assoc.*, **90**, 773–795.
- Liu,X. *et al.* (2001) Bioprospector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.*, **6**, 127–138.
- Makita,Y. *et al.* (2004) DBTBS: database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics. *Nucleic Acids Res.*, **32**, 75–77.
- McCue,L. *et al.* (2001) Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res.*, **29**, 774–782.
- McCue,L. *et al.* (2002) Factors influencing the identification of transcription factor binding sites by cross-species comparison. *Genome Res.*, **12**, 1523–1532.
- Moszer,I. *et al.* (1995) Subtilist: a relational database for the *Bacillus subtilis* genome. *Microbiology*, **141**, 261–268.
- Qin,Z.S. *et al.* (2003) Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites. *Nat. Biotechnol.*, **21**, 435–439.
- Remm,M. *et al.* (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.*, **314**, 1041–1052.
- Roth,F. *et al.* (1998) Finding DNA regulatory motifs within unaligned non-coding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.*, **16**, 939–945.
- Saxild,H. *et al.* (2001) Definition of the *Bacillus subtilis* PurR operator using genetic and bioinformatic tools and expansion of the PurR regulon with glyA, guaC, pbuG, xpt-pbuX, yqhZ-foID, and pbuO. *J. Bacteriol.*, **183**, 6175–6183.
- Shine,J. and Dalgarno,L. (1974) The 3'-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc. Natl Acad. Sci. USA*, **71**, 1342–1346.
- Wang,T. and Stormo,G. (2003) Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics*, **19**, 2369–2380.