BMC Bioinformatics

Research article

Open Access

# Genome wide identification of regulatory motifs in Bacillus subtilis

Michael M Mwangi and Eric D Siggia*

Address: Center for Studies in Physics and Biology, The Rockefeller University, 1230 York Avenue, New York, NY

Email: Michael M Mwangi - mmm37@cornell.edu; Eric D Siggia* - siggia@eds1.rockefeller.edu

* Corresponding author

Published: 16 May 2003

*BMC Bioinformatics* 2003, **4**:18

Received: 17 January 2003

Accepted: 16 May 2003

This article is available from: http://www.biomedcentral.com/1471-2105/4/18



## Abstract

**Background:** To explain the vastly different phenotypes exhibited by the same organism under different conditions, it is essential that we understand how the organism's genes are coordinately regulated. While there are many excellent tools for predicting sequences encoding proteins or RNA genes, few algorithms exist to predict regulatory sequences on a genome wide scale with no prior information.

**Results:** To identify motifs involved in the control of transcription, an algorithm was developed that searches upstream of operons for improbably frequent dimers. The algorithm was applied to the B. subtilis genome, which is predicted to encode for approximately 200 DNA binding proteins. The dimers found to be over-represented could be clustered into 317 distinct groups, each thought to represent a class of motifs uniquely recognized by some transcription factor. For each cluster of dimers, a representative weight matrix was derived and scored over the regions upstream of the operons to predict the sites recognized by the cluster's factor, and a putative regulon of the operons immediately downstream of the sites was inferred. The distribution in number of operons per predicted regulon is comparable to that for well characterized transcription factors. The most highly over-represented dimers matched $\sigma^A$, the T-box, and $\sigma^W$ sites. We have evidence to suggest that at least 52 of our clusters of dimers represent actual regulatory motifs, based on the groups' weight matrix matches to experimentally characterized sites, the functional similarity of the component operons of the groups' regulons, and the positional biases of the weight matrix matches. All predictions are assigned a significance value, and thresholds are set to avoid false positives. Where possible, we examine our false negatives, drawing examples from known regulatory motifs and regulons inferred from RNA expression data.

**Conclusions:** We have demonstrated that in the case of B. subtilis our algorithm allows for the genome wide identification of regulatory sites. As well as recovering known sites, we predict new sites of yet uncharacterized factors. Results can be viewed at http://www.physics.rockefeller.edu/~mwangi/.

## Background

Bacterial genome annotation has generally been confined to the prediction of sequences encoding proteins and prominent families of RNA genes. The predicted ORF's are grouped into categories by comparing them (e.g. using BLAST) to hand curated proteins or motifs with already characterized functions. Information about protein interactions can be extracted by finding how genes group into operons [1] and searching for homologs to protein

domains that reside on distinct proteins in one species and are joined into a single protein in another [2,3].

The comparison between the genomes of the model metazoans (fly, worm, and plant) with the human genome has confirmed the widely held belief of evolutionary and developmental biologists that much of the diversity of life stems from changes in regulation and not the creation of novel proteins. For bacteria, there is much more horizontal gene transfer, and it is an unresolved question of how regulation of these genes is coordinated with the host. When it is realized that even for E. coli less than 20% of the operons have been thoroughly examined upstream for regulatory motifs and less than 1/4 of the 300 or more putative DNA binding proteins have known sites, it is apparent that the automatic methods for inferring regulatory motifs must approach those used for inferring protein coding sequences and function if the full potential of the 'genomic revolution' is to be realized.

The inference of improbably frequent motifs from a collection of sequences is a recognized branch of bioinformatics. Algorithms can be categorized by the search strategy used to find a motif and the model used to assess the probability of the frequency of the motif's occurrence. Most algorithms operate on the regulatory sequenes of clusters of genes with related function and return one or a few motifs [4–10]. The probability of the frequency of occurrence of a given motif in the set of regulatory sequences is usually assessed based on the contrast between the set of sequences and the rest of the genome. When the genomes of other related species are available, motif predictions can by interspecies comparisons sometimes be made on a gene by gene basis [11–14]. Computational methods, though imperfect, are an essential step in interpreting genome wide experiments and doing a preliminary screen for targets that merit further laboratory investigation.

In this article, we extend a strategy originally applied to E. coli [15] that considers the entire genome at once and finds all improbably frequent motifs in parallel. It uses an exhaustive search, so it misses nothing within the category of motifs it searches for. Probability is assessed internally since there is no plausible set of sequences to compare against. This approach has the obvious merits of presuming nothing about regulons, being quick to implement, and using all the available sequences that may share regulatory motifs. It has the obvious demerits of not using the protein annotations or information about co-expression such as available from microarray experiments. Instead, these resources are used to check the validity of the putative regulons predicted from the sequence alone.

In bacteria, a regulatory protein often recognizes and binds to a class of similar dimers, where a dimer $W_1N_xW_2$ consists of two specific words $W_{1,2}$ separated by $x$ non-specific bases. If a dimer is observed to occur $n$ times in a set of sequences, a p-score can be assigned to the frequency $n$ by computing the probability of observing $n$ or more instances of the dimer under a null model that assumes the instances of the words $W_{1,2}$ are distributed in the sequences at random. For the p-score to be considered significant, it must fall below an appropriately chosen threshold. Since many regulatory proteins bind to dimers with identical or reverse complementary words $W_{1,2}$, these classes of dimers are given special consideration, and a different threshold is used than in the general case. Because secondary structure motifs also have dimer form, dimers with significant p-scores do not always represent protein binding sites, e.g. the T-box.

Essential to the success of our strategy is the way in which we cluster over-represented dimers, derive weight matrices, and infer regulons. Dimers are clustered into distinct groups based on sequence similarity. Weight matrices for the clusters are derived from the actual sequences matched by the dimers and scored over the regions upstream of the operons to predict sites. The set of operons immediately downstream of the matches to a particular weight matrix are inferred to be a regulon. We find in Results that the number of predicted regulons as well as the number of operons per predicted regulon is in line with expectations. For only a sixth of our clusters of dimers could evidence for function be deduced from the available information for B. subtilis. We validated these clusters by comparing their weight matrix matches with known regulatory sites, examining the operons composing their regulons for common function (either manually using the detailed gene annotations or automatically using the COG categories), and inspecting their weight matrix matches for positional biases (with respect to translation start or predicted $\sigma^A$ sites).

## Algorithm

Our algorithm for identifying regulatory elements in prokaryote genomes is an extension of [15] and consists of the following steps.

1. Identify operons and extract upstream sequences.

2. Enumerate statistically over-represented dimers of the form $W_1N_xW_2$ in upstream sequences.

3. Cluster the dimers into similar groups.

4. Construct for each cluster a weight matrix, derived from the matches in the upstream regions to the cluster's dimers.

5. Predict regulatory elements by using standard information theory to score the upstream sequences against the weight matrices.

For B. subtilis, the rate limiting steps 2 and 3 take ~1/2 and ~3 hr respectively to execute on a 500 MHz Pentium II workstation. Our predicted regulatory sites can be viewed at the URL http://www.physics.rockefeller.edu/~mwangi/.

### *Putative operons and upstream sequences*

We group adjacent ORF's on the same strand into putative operons if (a) the two ORF's are separated by no more than $m$ bases or (b) the two ORF's are both not hypothetical, are separated by no more than $n$ bases, and have names differing only in their last letters, which would suggest that the ORF's protein products have related functions. Tests involving E. coli K12 suggest that the optimal values of $m$ and $n$ are $m$ = 32 and $n$ = 130. At these values, we correctly predict ~70% of the ~400 operons in E. coli K12 listed in RegulonDB [16] as having some supporting experimental evidence. To construct the set of upstream sequences most likely to contain regulatory elements, we extract from immediately upstream of the translation start sites of our predicted operons a maximum of 300 bases, limited so as not to include any coding sequence (ORF) on either strand. The upper limit of 300 is chosen because it includes almost all the known regulatory sites in E. coli K12 [17]. Using the REPuter program [18], we discard all exact repeats of length 16 or more bases from the upstream contiguous sequences to eliminate potential insertion sequences and transposons. From the set of fragments generated by the removal of the repeats, we discard any fragment with less than 50 bases to obtain our final set of upstream sequences. Of the 471289 bases of upstream sequence in B. subtilis, 98.3% remained after the removal of the repeats, and 95.9% remained after imposing the minimum fragment length of 50.

### *Enumeration of dimers*

We search in the upstream sequences for statistically over-represented dimers of the form $W_1N_xW_2$ with word strings $W_1$ and $W_2$ of a, c, g, and t of lengths 4–5 and a spacing $x$ in the range 3–30. When we include words with length 3 or less, we find it virtually impossible to cluster the over-represented dimers, probably because dimers with short word lengths occur frequently in the whole genome and can be part of regulatory elements recognized by different transcription factors. When we use words with length 6 or greater, the large sample space of dimers searched for necessitates that we be exceedingly stringent with our thresholds for significance, so only the most improbably infrequent motifs are detected. Since the conserved portions of the consensus sequences of known regulatory elements are rarely observed to be separated by more than ~20 bases [19], it is natural to constrain $x$ to the interval 3–30. To enumerate the dimers, we tabulate the positions of all words $W$ in our set $U$ of upstream sequences in a three dimensional table, the entries of which are indexed by the string $W$ and the sequence $S$ in $U$ that contains the occurrence of $W$. We then use the table to count the number of occurrences $n(D)$ of the dimer $D = W_1N_xW_2$ in $U$. Denoting the length of a word or dimer $M$ as $L(M)$, the expected number of occurrences of $D$ under the null hypothesis that the occurrences of $W_1$ and $W_2$ are uncorrelated is

$$\gamma(D) = L_{eff}(D)\frac{n(W_1)}{L_{eff}(W_1)}\frac{n(W_2)}{L_{eff}(W_2)} \qquad (1)$$

where $n(W)$ is the total number of occurrences of a word $W$ in $U$ and $L_{eff}(M) = \Sigma_{S\in U}[L(S) - L(M) + 1]$ is the number of independent positions in $U$ that a motif $M$ can be placed. The probability $P$ of observing $n(D)$ or more occurrences of $D$ under our null hypothesis is then given by the Poisson distribution:

$$P = \sum_{n\geq n(D)} \frac{[\gamma(D)]^n}{n!} e^{-\gamma(D)} \qquad (2)$$

A dimer is considered over-represented if

$$P < 1/N_D \qquad (3)$$

where $N_D$ is the number of dimers considered, that is $(4^4 + 4^5)^2 \cdot 31 \sim 50{,}000{,}000$. Because the binding sites of transcription factors are frequently symmetric (e.g. acctN$_5$acct) or reverse complement symmetric (e.g. ccctN$_5$aggg) [19], we score these separately using $N_D = (4^4 + 4^5) \cdot 31 \sim 40{,}000$. Under our null model, no dimer would satisfy Eq. 3 by chance. However, our null model is inaccurate for sequences consisting of long stretches of the same nucleotide (A or T being the most common cases in practice) since a sequence like AAAAAAN$_x$TTTTTT can for $x < y$ contain multiple instances of the dimer AAAAN$_y$TTTT displaced relative to each other by one base. In contradiction to our null model, the occurrences of the words AAAA and TTTT are manifestly correlated, leading to extra instances of the dimer AAAAN$_y$TTTT and an over-estimation of the significance of the frequency of the dimer's occurrence. To circumvent the problem, we ignore all words that consist of only the same nucleotide and so miss motifs like AAAAN$_5$TTTT recognized by ComK. It is however important to note that the frequency of occurrence of a dimer like TAAAAN$_5$TTTTA is properly assessed since the problem is not the abundance of any particular

nucleotide in a dimer but the translational symmetry that results when each word is a continuous uninterrupted string of the same nucleotide. We therefore believe that our statistical model is misrepresenting a negligible number of motifs even in the poly A/T rich genome of B. subtilis.

### *Clustering of dimers*

Since many of our over-represented dimers represent different but overlapping versions of the conserved cores of the binding sites of the same factor, it is necessary for us to cluster our over-represented dimers into distinct groups. For example, the following two dimers

ttgaNNNNNNNNNNNNNNNNNNataat

tgccNNNNNNNNNNNNNNNNNtata

in B. subtilis should belong to the same group since they are both related to the consensus sequence $TTGACAN_{17}TATAAT$ recognized by the sigma factor $\sigma^A$ [20].

To cluster our dimers, we first compute for each pair of dimers $D_1$ and $D_2$ a pairwise similarity score $S(D_1, D_2)$. We define the score of an alignment of $D_1$ and $D_2$ to be a sum over the pairs of overlapping bases: matches are scored as +1, mismatches as -1, and N paired with anything scores as 0. We define $S(D_1, D_2)$ to be the maximum score produced by all possible alignments between $D_1$ and $D_2$ subject to the constraint that the left (and similarly right) words in the two dimers must partially overlap by at least 2 bases. When the aforementioned constraint cannot be satisfied, we define $S(D_1, D_2) = 0$. Hence, for the above two dimers, $S(D_1, D_2) = 5 - 1$.

Define a pairwise dis-similarity score between $D_1$ and $D_2$ as $D(D_1, D_2) = \max_{D_1, D_2} S(D_1, D_2) - S(D_1, D_2)$. Now, define a graph $G_0 = (V, E_0)$ with vertices $V$ representing our dimers and edges $E_0$ having lengths $D(D_1, D_2)$. In such a graph, highly similar dimers would tend to form spatially compact clusters. In practice, these compact clusters tend to be connected by long chains of edges since through a series of substitutions, insertions, and deletions a dimer $D$ can be transformed into a highly dis-similar dimer $D'$. Because of these long chains of edges, many clustering algorithms have difficulty properly delineating the compact clusters. For example, as the agglomerative algorithm CAST [21] constructs clusters starting with our individual dimers, it fails to merge many highly similar groups together, probably because the algorithm has a difficult time deciding which groups the dimers on the long chains of edges belong to. In the divisive SPC algorithm [22], our dimers are represented as spins in a Pott's model, and clusters are mapped out using a spin-spin correlation function. As the temperature is raised, the increase in thermal energy disrupts the correlation between many of the dimers in the compact clusters before it disrupts the correlations over the long chains of edges between the clusters, leading to many highly similar groups often each consisting of one dimer. We devised an algorithm, the weakest-link-clustering (WLC) algorithm, to specifically seek out and severe the long chains of edges to generate compact clusters.

Starting with $G_0 = (V, E_0)$, our WLC algorithm in each iteration generates from a graph $G_{i-1} = (V, E_{i-1})$ a new graph $G_i = (V, E_i)$. Our clusters of dimers are defined as the connected components of the current graph $G_i$. To generate $G_i$, we compute all the shortest finite paths in $G_{i-1}$ between all pairs of dimers $(D_1, D_2) \in V$. Multiple paths may run across a given edge $(D_1, D_2) \in E_{i-1}$. Let $P(D_1, D_2)$ be the mean length of these paths. The weakest link in $G_{i-1}$ is defined as the edge $(D_1, D_2) \in E_{i-1}$, $D_1 \neq D_2$, at which $P(D_1, D_2)$ is a maximum. When $P(D_1, D_2)$ is a maximum at multiple edges, then the edge to be designated the weakest link is chosen at random. Irrespective of the exact edge chosen to be the weakest link, the edge will undoubtedly be part of one of the aforementioned long chains of edges in $G_0$. To generate compact clusters of dimers, our algorithm severs the weakest link in $G_{i-1}$ to produce $G_i$.

Define the intra-cluster affinity $A(C)$ of a connected component $C = (V_c, E_c)$ as

$$A(C) = \frac{\sum_{D_1, D_2 \in V_c, (D_1, D_2) \in E_0} S(D_1, D_2)}{|V_c|^2} \qquad (4)$$

Note that the sum of similarity scores $S(D_1, D_2)$ in the numerator is performed over all pairs of dimers in $C$ (or equivalently edges in $E_0$) regardless of whether or not the dimers are currently connected by an edge in $C$. $A(C)$ is therefore the average pairwise similarity score of dimers in $C$. Every time two new connected components $C_1$ and $C_2$ are formed by severing a weakest link in a connected component $P$, our algorithm computes the ratio

$$R = \frac{[A(C_1) + A(C_2)]/2}{A(P)} \qquad (5)$$

of the mean of the infra-cluster affinities of the child clusters $C_1$ and $C_2$ to the intra-cluster affinity of the parent cluster $P$. Hence, our algorithm produces a series of $R$ values until the trivial state of every dimer being in its own cluster is reached.

The optimal number of clusters can be inferred from a plot of $R$ versus the number of clusters, e.g. Figure 1. $R$ declines rapidly as the highly non-compact clusters are severed and plateaus when a succession of clusters are encountered that exhibit the same degree of compactness from parent to children. $R$ can then increase, e.g. around 370 clusters in Figure 1, when formerly compact clusters are fragmented into yet better children. To make our clusters as generic as possible, we choose the cluster number to be the in the first plateau in $R$.

### *Weight matrices*

For each cluster of dimers generated by our WLC algorithm, we extract from our upstream sequences all unique segments that match any dimer. Using a multiple alignment of the dimers in the cluster, we align the segments and pad each to the left and right with up to ~5 bases from the genome to create a column of equal length segments. We compute a matrix $n_{i,\alpha}$ that gives the number of occurrences of the base α in the $i^{th}$ column of the alignment. We prune the matrix $n_{i,\alpha}$ by performing a chi-squared test over a window of length $l_c = 3$ columns running over the length of the matrix. For a given position of the window, we compute the probability [23] that the observed matrix entries were obtained by sampling the background distribution of frequencies $f_a^0, f_c^0, f_g^0$, and $f_t^0$ of the bases a, c, g, and t respectively in our upstream sequences. When the probability exceeds 1%, we block out the middle column in the window and do not use it to score sequences against the matrix. Although the outer low significance columns are eliminated from the matrix, the inner blocked out columns are retained to preserve the spacing. The final matrix typically has a dimeric pattern, but monomeric, trimeric, and even more complex patterns are occasionally observed.

To predict regulatory sites, we in accordance with a scoring scheme by Berg and von Hippel [24] first convert the pruned matrix $n_{i,\alpha}$ to a weight or surrogate binding energy matrix $w_{i,\alpha}$. For an unblocked column $i$, $w_{i,\alpha} = \log_{10}(f_{i,\alpha}/f_\alpha^0)$ where $f_{i,\alpha} = (n_{i,\alpha} + 1)/\Sigma_\alpha (n_{i,\alpha} + 1)$ is the relative frequency of the base alpha in the $i^{th}$ column with a pseudo count of 1 added due to the Bayesian estimate. For a blocked column $i$, $w_{i,\alpha} = 0$.

The consensus sequence for a weight matrix is computed according to the prescription outlined in [25]. Denote the total number of counts $\Sigma_\alpha n_{i,\alpha}$ recorded in the $i^{th}$ column of the matrix $n_{i,\alpha}$ by $N$. If $n_{i,\alpha}/N > 0.5$ for some base α and $n_{i,\alpha} > 2 \cdot n_{i,\beta}$ for all bases $\beta \neq \alpha$, then the $i^{th}$ site in the sequence is assigned the consensus α. Otherwise, if $(n_{i,\alpha} + n_{i,\beta})/N > 0.75$ for some pair of bases α and β, then the site is assigned the co-consensus [α/β]. If neither criterion is satisfied or if the column $i$ was blocked out because it did not satisfy the chi-squared test, the site is assigned a N.

### *Predicting regulatory sites*

The score of a sequence $b_1 b_2 ... b_l$ to a weight matrix $w_{i,\alpha}$ is defined by the sum $s = \sum_{i=1}^{l} w_{i,b_i}$, which correlates with the binding affinity of the factor to the DNA sequence [24]. When a weight matrix is scored over many distinct segments, the histogram of scores $s$ can usually be approximated by some normal distribution $N(s; m, \sigma)$ with mean $m$ and variance $\sigma^2$. Hence, we characterize a weight matrix by the mean $m_s$ and variance $\sigma_s^2$ of the scores of the matrix to the $N$ defining segments used to compute the matrix and by the mean $m$ and variance $\sigma^2$ of the scores of the matrix against all the distinct segments of length $l$ in our upstream sequences. The more separated $N(s; m_s, \sigma_s)$ and $N(s; m, \sigma)$ are, the better the matrix can distinguish potential sites from background sequences. The sites predicted by a weight matrix are those with a score larger than a cut-off $s_0$. The false positive rate is given by $\int_{s_0}^{\infty} N(s; m, \sigma)$ and the false negative rate by $\int_{-\infty}^{s_0} N(s; m_s, \sigma_s)$. Since a decrease in the false positive rate can only occur at the expense of an increase in the false negative rate, care must be taken in choosing $s_0$. We choose $s_0$ to be $\max\{m_s - z_{self}\sigma_s, m + z_{back}\sigma\}$, with the two parameters $z_{self}$ and $z_{back}$, called the critical self and background z-scores respectively, typically having the values 1 and 3 to ensure a false positive hit rate no greater than 0.2%.

### *Running time*

The rate limiting steps of our algorithm are the exhaustive search for and the clustering of the over-represented dimers. The exhaustive search executes in $O(N_D + L_U L_D)$ time where $N_D$ is the number of dimers searched for, $L_U$ is the combined length of the upstream sequences, and $L_D$ is the sum of the different word lengths considered (e.g. 9 if word lengths 4 and 5 are considerd). To reach the trivial state that every dimer is in its own cluster, our current implementation of our WLC algorithm executes in $O(|E|^2|V|\log_2|V| + |E||V|^3)$ time on the graph $G = (V, E)$ and uses a breadth first search to identify the connected components and Dijkstra's algorithm to compute the shortest paths between all over-represented dimers. The stated running time of our WLC algorithm however should be interpreted as an upperbound that can in certain instances be a gross overestimate depending on the precise topology of the graph $G$.

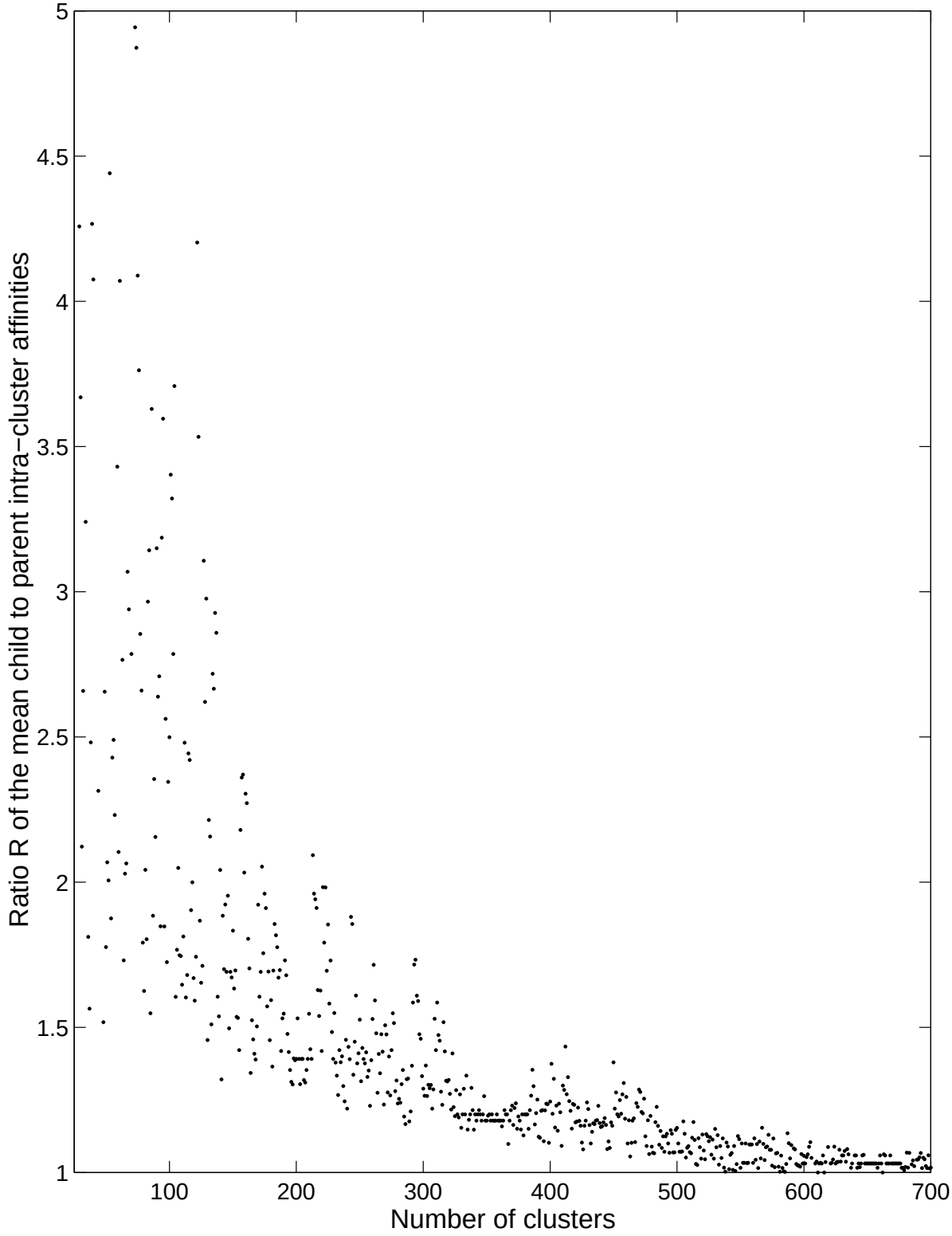


**Figure 1**
WLC Algorithm. Choosing the correct number of clusters. The ratio R (Eq. 5) of the mean child to parent infra-cluster affinities versus the number of clusters for B. subtilis generated by our WLC algorithm. As weakest links are severed, the number of clusters increases from 29 to 732. Note the stabilization at and around 350 clusters, the optimal cluster number.

**Table 1: The top 10 most significant dimers (column 1). Dimers searched for had word lengths 4–5 and a spacer 3–30. Coding sequence was not considered. Listed are the number of occurrences in the dataset (column 2) and the statistical significance $-log_{10} P$ (column 3), with $P$ calculated from Eq. 2.**

| Dimer | Number observed | Significance |
|---|---|---|
| ttga$N_{20}$ataat | 48 | 21.1 |
| gccgc$N_{11}$gcggc | 10 | 15.9 |
| ggtgg$N_3$cgcg | 10 | 14.6 |
| ttga$N_{19}$tata | 70 | 14.5 |
| gaaac$N_{16}$Cgta | 17 | 14.4 |
| ttga$N_{21}$taat | 58 | 13.8 |
| agggt$N_4$ccgcg | 8 | 13.7 |
| gccgc$N_{12}$Cggc | 12 | 13.7 |
| ttga$N_{23}$ataa | 75 | 13.6 |
| ttgac$N_{19}$ataat | 18 | 13.6 |

### *Web site*

Our website http://www.physics.rockefeller.edu/~mwangi/ presently lists our regulatory site predictions for B. subtilis and several other species. For each, we list

1. the regulon defined by each weight matrix with the operons' annotations,

2. each upstream region with the predicted regulatory sites,

3. all matrices with significant number of multiple matches in a upstream region,

4. pairs of our matrices that frequently co-occur in the same upstream region,

5. the observed and expected distributions of positions of a matrix's matches relative to translation start,

6. the observed and expected distributions of positions of a matrix's matches relative to our best primary sigma factor binding site predictions,

7. input files for the DNA sequence viewer and annotation tool ARTEMIS [26].

## Results

### *Nomenclature*

To simplify terminology, we will use the term 'operon' in what follows to denote our putative operons predicted as described in "Putative operons and upstream sequences." Since a particular weight matrix is thought to represent sites uniquely recognized by some transcription factor, the term 'regulon' will be used for the group of operons having a match to the matrix directly upstream, i.e. direct targets of the factor.

### *Overview*

We applied our algorithm to the well studied gram positive bacteria B. subtilis. We grouped the 4100 prokaryote's ORF's into 2729 putative operons and found after the removal of poly a/t patterns 732 over-represented dimers with both words between 4–5 in length and a spacer between 3 and 30. In the list of our 10 most significant dimers in Table 1, the four dimers ttga$N_{20}$ataat, ttga$N_{19}$tata, ttga$N_{21}$taat, and ttgac$N_{19}$ataat all correspond to the consensus sequence TTGACA$N_{17}$TATAAT recognized by the primary sigma factor $\sigma^A$ [20], the two dimers ggtgg$N_3$cgcg and agggt$N_4$ccgcg correspond to the T-box [27] with a known consensus sequence AANNAGGGT-GGTACCGCGNN involved in the alternate transcription termination regulation of the aminoacyl-tRNA synthetases, and the consensus sequence TGAAACN$_{16}$CGTA recognized by the antimicrobial resistance sigma factor $\sigma^W$ [20] is represented by the dimer gaaac$N_{16}$cgta.

Figure 1 shows the ratio $R$ (Eq. 5) of the mean child to parent intra-cluster affinity versus the number of clusters when we clustered the 732 over-represented dimers. There was a plausible plateau at 350 clusters, 97 of which contained 2 or more dimers. We found that 317 of the 350 clusters matched 3 or more sequences and converted these clusters to weight matrices for further study. Of the 317 matrices, we were able to identify 52, listed in Table 2, that met at least one of our criteria for significance. Of the 52 matrices, 10 represent experimentally characterized regulatory factors, 30 have regulons that contain a disproportionate number of operons with related functions, and 32 have matches exhibiting some positional bias. A total of 28 of the matrices listed in Table 2 were derived from a cluster of two or more dimers. To further demonstrate our algorithm, we also searched for longer symmetric dimers with word lengths 6 that could overlap coding sequence and applied our algorithm to the subset of sequences

**Table 2: 52 unique biologically significant weight matrices. Listed are the matrix's identifier (column 1), consensus sequence (column 2), regulon size (column 3), and annotation (column 4). The matrices are sub-divided into categories according to the means by which they were identified: by comparison to documented regulatory mechanisms, by inspecting the operons in a matrix's regulon for related functions, and by examining the matrix's matches for positional biases. If a matrix was identified by several means, all listings for the matrix except the first in the top-most category are marked with pluses. Where applicable, the statistical significance – $\log_{10}$ P is reported in (), and entries in a category are sorted according to significance.**

| Weight matrix | Consensus sequence | Regulon size | Annotation |
|---|---|---|---|
| | **Documented regulatory mechanisms** | | |
| | ***DBTBS database***[28] | | |
| | *Sigma factors* [20] | | |
| WM1 | $N_7$TTGAN$_{19}$TATAATAN$_6$ | 1141 | $\sigma^A$, housekeeping |
| WM1118 | [G/T]GTTTAN$_{13}$ [A/C]GGGAA [G/T] | 8 | $\sigma^B$, general stress response |
| WM11 | NTGAAACNTTTN$_{12}$CGTAT [A/T] | 16 | $\sigma^W$, antimicrobial resistance |
| WM212 | TGGCA [C/T]N$_4$CTTGCAT | 5 | $\sigma^{L1}$, levanase and amino acid catabolism |
| | *Miscellanous* | | |
| WM2 | AANNAGGGTGGTACCGCGNN | 24 | T-box, alternate transcription termination regulation of aminoacyl-tRNA synthetases [27] |
| WM22 | [A/T]AAN [A/C]GAACNN [A/T]NGTTCNNTTN | 29 | LexA, SOS response [36] |
| WM71 | NT [A/T]TGTAN$_{10}$ACA [A/T]AN | 111 | TnrA, pleiotropic regulator involved in global nitrogen regulation [37] |
| WM317 | [A/T]TGTAA [A/G]CG [C/T]TT [A/T]N [A/T] | 54 | CcpA, carbon catabolite repression [59] |
| | ***Two-component response regulators*** [43,44] | | |
| WM298 | NTAATN$_{20}$ATTAN | 27 | YccG-YccH (3.4) |
| WM259 | TGCGN$_{10}$CGCA | 5 | YclK-YclJ (3.3) |
| | **Novel predictions** | | |
| | ***Regulons which operons have highly related functions*** | | |
| | *Identified by detailed manual inspection* | | |
| WM171 | TGGGN$_{11}$GGGA | 2 | Sec-dependent protein export machinery |
| WM116 | AATTC [A/T]N$_{28}$ [A/T]GAATT | 4 | Cell lysis |
| WM266 | TGGACAN$_3$GCAGA | 3 | Extracellular proteins |
| WM304 | AGTGTN$_{15}$AGACT | 4 | Transport |
| WM69 | TATCTN$_4$ [A/T]TCGAGA | 5 | Transport |
| WM233 | NGGGAN$_3$TGCGG | 7 | Antimicrobial resistance |
| WM290 | NTTGAN$_{16}$TGTTAN$_3$T | 18 | DNA synthesis and repair |
| WM47 | A [A/T]AGAGN$_{18}$CTCTTT [C/T]N | 27 | DNA synthesis and repair |
| WM124 | NTTAG [A/T]N$_6$TTAGN | 17 | Transport |
| | *Identified using COG functional categories* | | |
| +WM2 | AANNAGGGTGGTAGGGCGNN | 24 | T-box, translation, ribosomal structure, and biogenesis (12) |
| +WM317 | [A/T]TGTAA [A/G]GG [C/T]TT [A/T]N [A/T] | 54 | CcpA, carbohydrate transport and metabolism (6.5), energy production and conversion (3.0) |
| WM130 | N$_4$TTGAN$_{14}$ [A/T]N$_4$TGAAAN | 38 | Posttranslational modification, protein turnover, and chaperones (4.2) |
| +WM1 | $N_7$TTGAN$_{19}$TATAATAN$_6$ | 1141 | $\sigma^A$, transcription (3.3) |
| +WM212 | TGGCA [C/T]N$_4$GTTGCAT | 5 | $\sigma^L$, energy production and conversion (3.1) |
| WM255 | NCTGAAN$_{26}$TTCAGN | 3 | Cell motility and secretion (2.9) |
| +WM22 | [A/T]AAN [A/C]GAACNN [A/T]NGTTCNNTTN | 29 | LexA, DNA replication, recombination, and repair (2.6) |
| WM39 | [A/G]NNTGCTN$_{30}$AGCAN | 21 | Secondary metabolites biosynthesis transport, and catabolism (2.5) |
| WM228 | NGCAGAN$_{13}$TCTGCN | 3 | Secondary metabolites biosynthesis transport, and catabolism (2.5) |
| WM283 | AGCTGN$_{13}$GAGGTT | 3 | Translation, ribosomal structure, and biogenesis (2.4) |
| WM80 | NGTTTN$_{29}$AAACN | 86 | Energy production and conversion (2.3) |
| WM223 | NATTTN$_{28}$AAATN | 69 | Transcription (2.3) |
| WM16 | NCCGGC [C/T]N$_6$GCCGGN [G/T]TTTT | 27 | Signal transduction mechanisms (2.3) |
| WM17 | [A/G]NCGGCN$_8$ [A/G]NGCCGN | 40 | Cell motility and secretion (2.3) |
| WM23 | [A/T]CGAAN$_{27}$TTCG [A/T] | 25 | Amino acid transport and metabolism (2.2) |
| WM221 | NGCGGN$_{29}$CGGCN | 6 | Amino acid transport and metabolism (2.2) |
| WM119 | NAATAN$_9$TATTN | 62 | Cell envelope biogenesis, outer membrane (2.1) |
| +WM304 | AGTGTN$_{15}$ACACT | 4 | Inorganic ion transport and metabolism (2.1) |
| WM46 | NTATAN$_{17}$AAAGGAG [A/G]N | 109 | DNA replication, recombinaion, and repair (2.1) |
| WM75 | [G/T]N$_3$CTACN$_9$GN$_{12}$CTACA | 5 | Secondary metabolites biosynthesis transport, and catabolism (2.0) |
| WM31 | NTGTTN$_5$AACAN | 58 | Carbohydrate transport and metabolism (2.0) |
| | ***Positions of binding sites are highly biased with respect to $\sigma^A$ sites.*** | | |
| +WM46 | NTATAN$_{17}$AAAGGAG [A/G]N | 109 | Repressor (17) |
| WM21 | AANGCGN$_{15}$GGGNTTTTTT | 128 | Activator (7.9) |
| WM33 | NAAGC [A/T]GN$_{12}$C [A/T]GCTTN | 96 | Activator (4.7) |
| WM50 | NNGGTTTTTTATTN | 152 | Activator (3.6) |
| WM173 | NAAAGN [A/G]NGGAAN$_4$ | 35 | Repressor (3.0) |
| WM169 | NAAAGN$_3$GTGAN | 40 | Repressor (2.9) |
| WM13 | [A/G] [A/C] [A/G]CGG [G/T]... [G/T]N$_9$GGG [G/T] [G/T]TT [A/T]T | 21 | Activator (2.8) |
| WM180 | [A/T]AGAGN$_5$AGAGN | 15 | Repressor (2.6) |
| WM58 | NAAAGANAN$_{15}$TGTTTTN | 42 | Activator (2.6) |
| WM79 | NTTGT[A/T N$_4$TTGTN | 67 | Activator (2.5) |

**Table 2: 52 unique biologically significant weight matrices. Listed are the matrix's identifier (column 1), consensus sequence (column 2), regulon size (column 3), and annotation (column 4). The matrices are sub-divided into categories according to the means by which they were identified: by comparison to documented regulatory mechanisms, by inspecting the operons in a matrix's regulon for related functions, and by examining the matrix's matches for positional biases. If a matrix was identified by several means, all listings for the matrix except the first in the top-most category are marked with pluses. Where applicable, the statistical significance – $\log_{10}$ P is reported in (), and entries in a category are sorted according to significance.** ***(Continued)***

| | | | |
|---|---|---|---|
| WM84 | $AN_3AACATN_3GGAGGN$ | 19 | Repressor (2.4) |
| WM7 | $NAAAGN_{19}$ [G/T]$CTTTN_3$ | 90 | Activator (2.3) |
| +WM17 | [A/G]$NGGGGN_8$ [A/C]NGCCGN | 40 | Activator (2.1) |
| | ***Absolute positions of binding sites are highly biased.*** | | |
| +WM46 | NTATA-17-AAAGGAG [A/G]N | 109 | (61) |
| +WM1 | $N_7TTGAN_{19}TATAATAN_6$ | 1141 | $\sigma^A$ (16) |
| +WM169 | $NAAAGN_3GTGAN$ | 40 | (10) |
| +WM21 | $AANCCGN_{15}CGGNTTTTTT$ | 128 | (6.3) |
| +WM2 | AANNAGGGTGGTAGGGGGNN | 24 | T-box (4.8) |
| +WM16 | NCGGGG [C/T]-6-GGCGGN [G/T]TTTT | 27 | (4.1) |
| +WM13 | [A/G] [A/C] [A/G]CCC[G/T ... | 21 | (3.9) |
| +WM58 | NAAAGANA-15-TGTTTTN | 42 | (3.4) |
| +WM11 | $NTGAAACNTTTN_{12}CGTAT$ [A/T] | 16 | $\sigma^w$ (3.1) |
| +WM17 | [A/G]$NCGGCN_8$ [A/C]NGCCGN | 40 | (3.0) |
| WM25 | NNGTTT-17-GG [A/T]A [A/T] | 59 | (3.0) |
| WM37 | NAAGC [A/T]-19-GCTTT | 25 | (3.0) |
| WM14 | $N_3CGGCN_{11}GCCGN_3$ | 197 | Tends to co-occur with T-box (3.0) |
| WM143 | $NCGTCN_{24}TTATN$ | 25 | (2.8) |
| WM185 | NAACC-15-GGTTNNTT | 15 | (2.7) |
| +WM47 | A [A/T]$AGAGN_{18}CTCTTT$ [C/T]N | 27 | (2.6) |
| +WM33 | NAAGG [A/T]$GN_{12}C$ [A/T]GCTTN | 96 | (2.1) |
| WM28 | [A/G]AAAGC-21- [A/G]GCTT [C/T]TT | 30 | (2.0) |
| | ***Unusually high number of matches in a single promter.*** | | |
| WM34 | NCACA [A/T]N [A/T]TGTGN | 17 | Three repeats overlap dnaA boxes TTATCCAGA [60], may inhibit chromosome replication, (7.8) |

upstream of operons identified to be co-expressed in various studies.

### *Regulon sizes*

To validate our methods, we began with a collection of transcription factors with experimentally characterized recognition sites collected in the DBTBS database [28] and by Helmann [29]. We restricted our attention to the 34 factors each with at least two sites, giving 600 sites in total. In the histograms of regulon sizes in Figure 2, size is reported in terms of number of operons. In (a), a regulon of a factor is defined as the set of our predicted operons that have immediately upstream a site documented in the DBTBS database to be recognized by the factor. Similarly, a regulon of a weight matrix in (b) or (c) consists of the operons that have immediately upstream a match to the matrix. A matrix in (b) for a factor was computed from the sites listed in the DBTBS database for the factor. For factors like $\sigma^A$ and DegU that recognize dimers with variable spacing *x*, we computed separately a matrix for each spacing *x*. The matrices in (c) are our 317 matrices derived from our clusters of over-represented dimers. As noted in the caption, a number of matrices have regulons containing more than 400 operons. Some of these matrices, like those for $\sigma^G$ and $\sigma^K$, were derived from experimental sites exhibiting little consensus. Others represent factors like ComK that recognize ubiquitous motifs like $AAAAN_5TTTT$, which may not all be functional. An exception is the matrix for the factor SpoOA with 824 operons in its regulon. SpoOA is the master regulator of sporulation and may have many targets [30].

The expressions in "Predicting regulatory sites" for the rates of false negative and positive predictions for a weight matrix's matches assumed a Gaussian distribution of values. We used the 600 DBTBS sites to test the expressions by computing for each factor's weight matrix in Figure 2(b) the percentage of its sites and the percentage of sites annotated for another factor that the weight matrix matched. The false negative rate can be deduced from the former percentage, and the false positive rate is given directly by the latter. The results agree well with the Gaussian assumption.

More than half of the regulons for our matrices in Figure 2(c) contain 10 or fewer operons. The five largest regulons with sizes 1141, 903, 518, 320, and 281 belong to the matrices WM1, WM5, WM29, WM4, and WM90 with consensus sequences $N_7TTGAN_{19}TATAATAN_6$,$N_5$[A/T]TTT [A/T]$N_5AAAT$[A/T][A/T]$N_5$,$NAAATTAN_6$[A/T]$N_4TAATTT$ NN,$N_4$[A/T]AAATT[A/T]$N_6$A[A/T]TT[A/T]$N_5$, and N[C/T]TTAC[A/T]$N_{16}GTAA$[A/G]NN respectively. Since WM1 represents the primary sigma factor $\sigma^A$, it is not surprising that its regulon contains nearly half of all our predicted

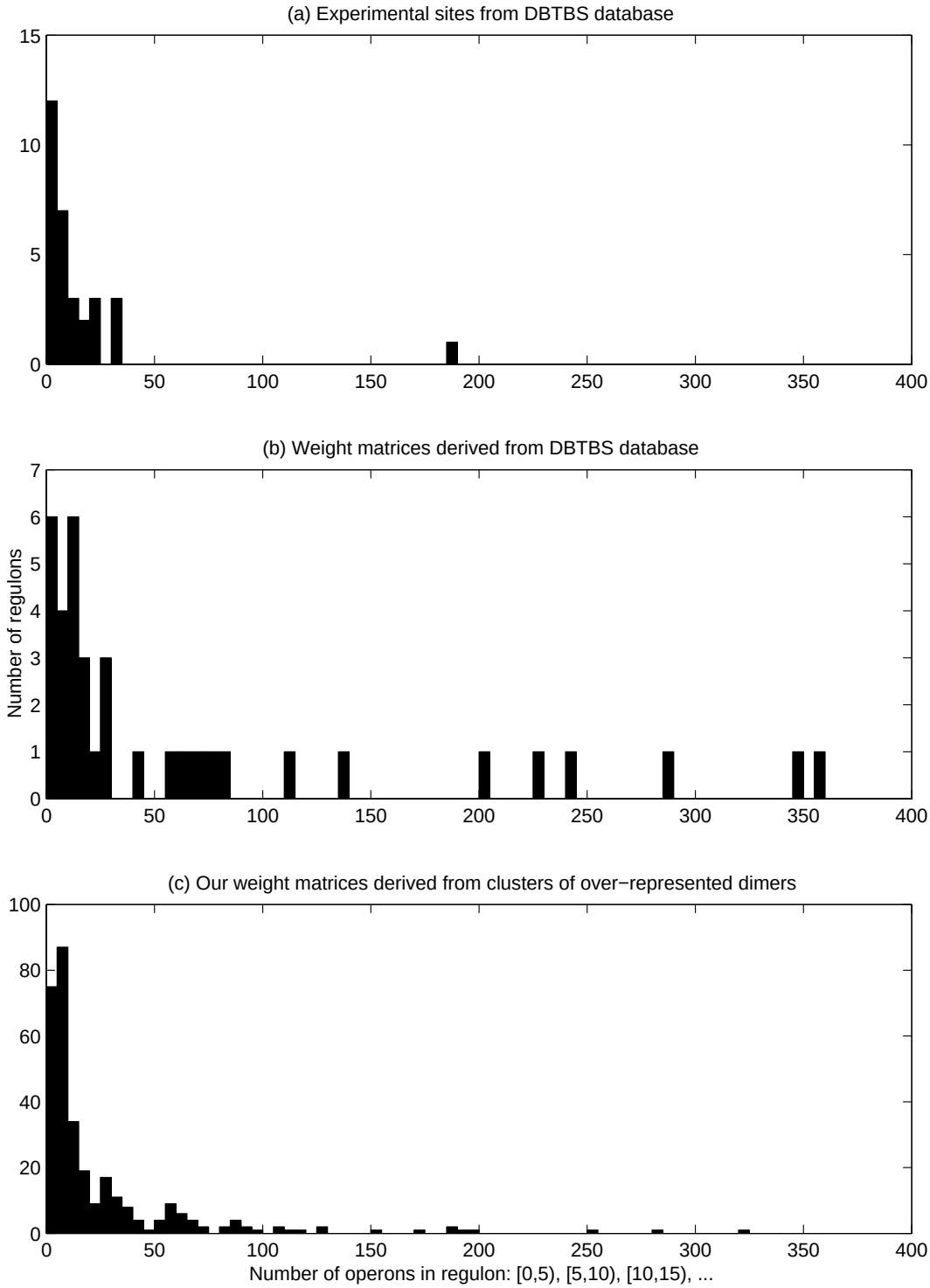


**Figure 2**
Histogram of regulon sizes. A regulon for a factor in (a) is defined as the set of our predicted operons that have immediately upstream a site documented in the DBTBS database to be recognized by the factor. A regulon for a weight matrix in (b) and (c) is defined as the set of our predicted operons that have immediately upstream a match to the matrix. The matrices in (b) were derived from the experimental verified sites in the DBTBS database. The matrices in (c) were derived from our clusters of over-represented dimers. The several regulons in (b) ($\sigma^A$, $\sigma^G$, $\sigma^K$, ComK, GltC, GltR, Hpr, LevR, and SpoOA) and the three regulons in (c) with more than 400 members are discussed further in the text.

operons. The matrices WM5, WM29, and WM4, representing ubiquitous poly a/t patterns, may correspond to UAS and UP elements [31]. The matrices that correlate well with the known factors $\sigma^B$, $\sigma^W$, T-box, $\sigma^L$, LexA, TnrA, and CcpA (see next section) have regulons containing 8, 16, 24, 5, 29, 111, and 54 of our operons respectively. It is clear from our literature search that we underestimate the number of operons directly targeted by $\sigma^B$ and $\sigma^W$. To date, at least 35 operons have been shown experimentally to be transcribed from $\sigma^B$ dependent promoters [32]. Moreover, various genetic and reverse genetic approaches and array technologies suggest that over 200 genes are $\sigma^B$ dependent, although some indirectly [33]. Using consensus search, run-off transcription followed by macroarray analysis, and transcriptional profiling, [34] identified 30 $\sigma^W$ dependent promoters. Our underestimates of the $\sigma^B$ and $\sigma^w$ regulons can be attributed to the high specificity of our weight matrices and highlight a weakness in our algorithm. Our $\sigma^B$ matrix was derived from a cluster of only one dimer, and our $\sigma^W$ matrix was derived from a cluster of 7 dimers with no mismatches. Although the factors' recognition consensuses are very well reflected by the dimers in these clusters, no dimers representing allowed variations to these consensuses in both sequence and spacer met our criterion for over-representation, so our $\sigma^B$ and $\sigma^w$ clusters and hence their derivative matrices were too specific and matched only the strongest of sites. To remedy this weakness, we would have to search not for over-represented dimers but over-represented classes of dimers with mismatches and variable spacers. Notwithstanding this, the sizes of our other matrix regulons compare favorably with those documented in the literature. Of the 21 aminoacyl-tRNA synthetase operons, 14 are known to be regulated by the T-box [35]. Reference [20] estimates that the $\sigma^L$ regulon contains 6 operons, and according to [36], some 20 operons are direct targets of LexA. We could not find any recent estimates of the sizes of the TnrA and CcpA regulons. Both factors are believed to regulate many genes [37,38], and CcpA according to the DBTBS database is believed to directly target at least 34 sites. Excluding WM5, WM29, and WM4, our 317 matrices predict on average 3.5 sites per upstream region. On our web site, we mark simultaneously all predictions from our 317 matrices and the matrices derived from the experimental sites.

### *Weight matrices correlating with known factors*

To correlate our 317 weight matrices with known factors; we scored them over the 600 DBTBS and Helmann sites. The number $e$ of sites for a factor $f$ expected to match a matrix $w$ is

$$e = \sum_{i=1}^{N_s} (l_i - l_w + 1) p_w \qquad (6)$$

where $N_s$ is the number of sites listed for the factor, $l_i$ is the length of the $i^{th}$ site $s_i$, $l_w$ is the number of columns in $w$, and $p_w$ is the probability that a randomly chosen segment of length $l_w$ will match $w$. The probability $P(w, f)$ of observing by chance that $n$ of the sites $s_i$ bound by $f$ contain a segment that matches $w$ is then given by the Poisson distribution. For $P(w,f)$ to be significant, we insisted that it be less than $10^{-4}$, roughly the inverse of the total number of matrices times factors being compared.

If $P(w, f)$ happened to meet this cutoff, we manually checked that it was the most conserved positions in the experimental sites for a factor that matched our matrix. By these criteria, we could correlate at least one of our matrices with one of seven factors ($\sigma^A$, $\sigma^B$, $\sigma^W$, $\sigma^L$, LexA, TnrA, and CcpA) in our collection of 34 factors, a 21% success rate. We missed the factors ComA, CtsR, Fnr, HrcA, RocR, YqhN, and Zur even though these factors are quite specific. In most cases, it is because the dimers that are part of the conserved core of the binding sites of the factor did not satisfy criterion Eq. 3. When we only considered dimers with word lengths 4, the number of dimers considered decreased ∼16-fold in the general case (∼4-fold in the symmetric cases), and we found that under the new less stringent criteria given by Eq. 3 that at least one of our matrices correlated with one of nine factors ($\sigma^A$, $\sigma^B$, $\sigma^W$, LexA, CcpA, ComA, Fnr, GltC, and GltR) or ∼26% of the factors in our collection. Unfortunately, we missed the factors $\sigma^L$ and TnrA since their representative dimers all contain length 5 words.

For $\sigma^A$, we can compare our predictions with those of Reference [39]. Using a hidden Markov model (HMM) fitted to known $\sigma^A$ sites that allowed for variable spacing between the -35 and -10 elements, [39] predicted 881 $\sigma^A$ sites in our upstream sequences, 625 with a spacer 17 captured by our weight matrix WM1. Our matrix WM1 has 1580 matches upstream of 1141 operons, of which 413 agree with one of the 625 from [39]. Moreover, WM1 matched with no training 109 out of the 132 sites listed in [29]. It is unclear if we are seriously over-predicting since [39] estimates that their HMM misses 30% of real sites, and some of our WM1 matches could represent other spacings, which would be expected to yield a disproportionately large number of false positives. Our prediction that ∼40% of our operons directly depend on the dominant sigma factor does not seem excessive. Our WM1 matches also have a very strong positional bias (see below).

Noticeably absent from our list are matrices that represent the very specific HrcA and Fur factors. HrcA binds to the CIRCE elements TTAGCACTCN$_9$GAGTGCTAA [40] directly upstream of the genes hrcA and groES. Although Fur recognizes the 15 bp consensus TGAtAATNATTaTCA,

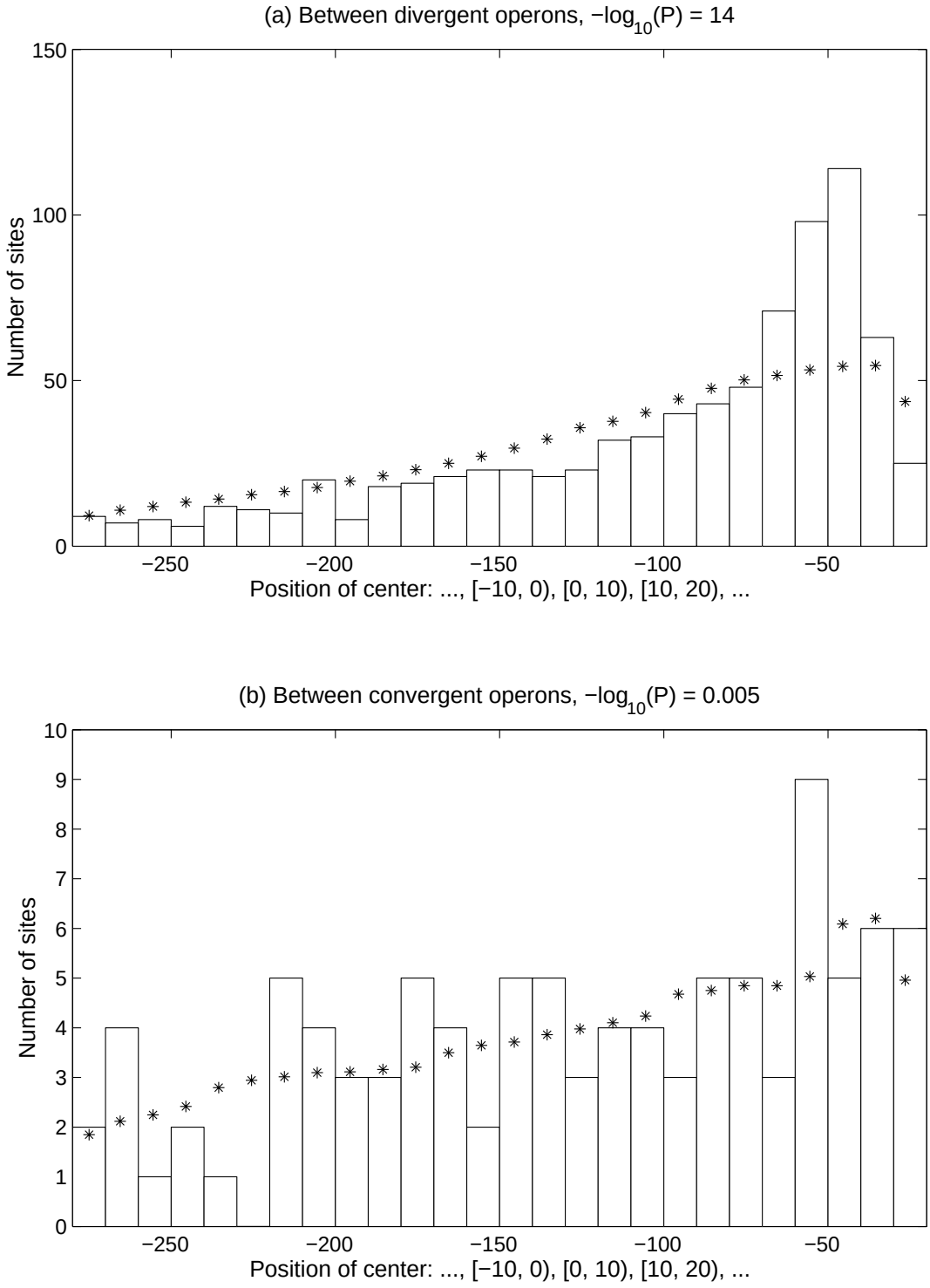


**Figure 3**
Matches to our $\sigma^A$ weight matrix WM1 exhibit a clear positional bias. Histograms of positions of the matches to our $\sigma^A$ weight matrix WM1 between (a) divergent and (b) convergent operons. In (a), positions are measured relative to translation start. In (b), positions are measured relative to the downstream end of the region. In either case, the first upstream base is assigned the position -1. The expected distribution, under the null hypothesis that the matches are uniformly distributed in their upstream regions, is denoted by *. Probability *P* of the observed distribution under the null hypothesis is reported as the significance score $-log_{10}P$.

many of the 20 operons known to be targeted by Fur are regulated by two overlapping Fur sites with the classic 19 bp consensus GATAATGATNATCATTATC [41]. One of the two CIRCE elements (that upstream of the groE operon) overlaps coding sequence and so was not in our search space of upstream sequences. We examined all dimers with word lengths 4–5 and spacing 1–11 that matched the given 15 and 19 bp Fur consensuses. The most significant dimer taatNNttatc occurred 15 times in our search space of upstream sequences with a probability $-log_{10}P = 3.6$. Although the Fur sites display a high degree of conservation along their length, it appears that due to variations at individual sites no dimer met our criterion for over-representation. For example, the dimer tgata$N_5$tatca only occurs twice in the 21 known Fur sites listed in Figure 4 in [41] because of variations at the fourth and twelfth sites highlighted in lowercase in the consensus TGAtAATNAT-TaTCA. This illustrates a known shortcoming of our method, which ignores dimers that, though not significant individually, are significant as a cluster. Because words like taat and ttatc occur frequently, note that the significance of the occurrence of the dimer taatNNttatc is much lower than might be expected. Under a null model ignoring nucleotide correlations, the dimer taatNNttatc would in our search space of ~0.5 Mb be expected to be seen $\sim 1/4^9 \cdot 500000 \sim 2$ times. Hence, the probability of seeing the dimer 15 times would by Poisson statistics be $\sim 2^{15}/15! \cdot e^{-2} \sim 10^{-9}$.

Observing that both the HrcA and Fur consensuses are long and reverse complement symmetric, we decided to search for over-represented symmetric and reverse complement symmetric dimers with word lengths 6 and spacers 1–30. We also augmented our search space to 300 bp upstream of each operon irrespective of whether this includes coding sequence. We clustered the 64 dimers we found into 35 clusters. Our third largest cluster had two dimers gcactc$N_9$gagtgc and tagcac$N_{13}$gtgcta that matched the CIRCE element and an additional two dimers acacgc$N_7$gcgtg and aagctc$N_{13}$gagctt that may define a broader recognition consensus for HrcA. There were no plausible matches to the Fur consensuses.

### *Known sets of coregulated genes*

Drawing from microarray studies, known regulons, CHip-CHip studies, etc., we compiled 39 sets, each containing genes believed to be targeted by some factor either directly or indirectly. The list of factors considered includes seven sigma factors (D, E, F, G, K, and X [20] and H [42]), the two-component systems DegU, ComA, and PhoP [43], 24 other two-component systems [44], AbrB [28], Fur [41], PucR [45], and PurR [46]. In what follows, we work with our operons predicted as discussed in "Putative operons and upstream sequences." If in a set a gene in one of our operons is listed to be a target of a factor, then the entire operon is considered to be a target.

For each of our 39 sets except those for the sigma factors, we asked which if any of our weight matrices had a regulon containing a disproportionate number of the set's operons. A weight matrix identified in this way could correspond either to the master factor believed to co-regulate the set's genes or some downstream factor activated later in the regulatory cascade. The probability that one of our regulons and one of the 32 sets considered share $n$ operons by chance can be assessed using Poisson statistics. A probability cutoff of $< 10^{-4}$ is used appropriate to our sample size of $32 \cdot 317$. None of our weight matrices were found to have a regulon containing a disproportionate number of the targets listed in a set. (Of the 3 (18) operons identified as being regulated by the two-component system YccG-YccH (YclK-YclJ), 2 (2) were contained in the regulon for our matrix WM298 (WM259) with a size 27 (5) for a significance of $\sim 10^{-3}$ ($\sim 10^{-3}$). We report these two-component systems in Table 2 because we think they might be real.)

To see why we did so poorly, we examined in more detail the 69 operons listed in the microarray study [43] to be targeted by DegU. Only 2 of these operons could be found among the 13 listed in the DBTBS database to be part of the DegU regulon. This suggests that the microarray study produced a considerable number of false negatives. It is possible that many of the targets listed in the study are indirect targets controlled by a cascade of regulatory factors and that no single factor directly binds to enough sites for its recognition consensus to be identifiable in our whole genome wide analysis.

For 13 of the factors we considered (sigma D, E, F, G, X, and H, DegU, PhoP, AbrB, Fur, PucR, and PurR), the recognition consensuses are known. For each factor, we applied our algorithm to the regions upstream of the operons listed in our sets to be coregulated by the factor. Since for each of these factors only 5–70 operons are listed to be targets, we had to search for only length 2–3 words in order to have reasonable counts. For 2/5 of the factors, at least one of our three topmost significant dimers matched the known consensus. For gene sets this small, other methods however may be preferable (see Discussion).

### *Regulons identified by operon functions*

A detailed manual examination reveals that the constituent operons of many of the regulons of our 317 weight matrices have highly related functions. For instance, the two operons prsA and sipA in the regulon for WM171 are both part of the Sec -dependent protein export machinery [47]. In addition, the regulon for WM304 of size four con-

tains at least three transporters, and the regulons for WM290 and WM47 contain a disproportionate number of genes involved in DNA synthesis and repair.

To attach a putative function to our 317 matrices automatically, we made use of ~20 COG functional categories assigned to the ORF's in the protein table (PTT) file for B. subtilis [48]. We defined the category of one of our operons to be the category of the first gene and inspected each of our matrix regulons for over-represented categories by using Poisson statistics to assess the number of operons belonging to any category. For significance, we used a probability threshold of 0.01, roughly the inverse of the number of regulons considered. (The probability threshold 0.001, corresponding to the inverse of the number of regulon-category pairs considered, would be too stringent since the over-representation of a particular category in one of our regulons often excludes the over-representation of another category.) The 21 of our matrices whose regulons contain an over-represented category are listed in Table 2 along with their significance scores. For a category to be over-represented in a given matrix regulon, note that the majority of the operons in the regulon need not belong to the category, just a disproportionately large number. Since many of the genes in B. subtilis have yet to be assigned a COG and since many regulons might contain operons belonging to a diverse set of categories, this form of automatic functional scoring is rather haphazard. Indeed, only one of the matrices WM304 that we identified manually (using the more extensive information available at http://genolist.pasteur.fr/SubtiList/[49]) came up in our automatic screen. When we searched the regulons consisting of our operons immediately downstream of the experimental verified sites listed in the DBTBS database and by Helmann, the regulons for 5 out of the 11 sigma factors and 6 out of the 23 other transcription factors contained an over-represented category.

### *Weight matrices with positional bias*

When we score our matrices over the whole genome, the matches to some of our matrices exhibit clear positional biases. Not only do many of these matches prefer to fall in non-coding as opposed to coding sequence, which can be expected since the matrices themselves are derived from non-coding sequence, but the matches tend to cluster into various intervals upstream of the translation starts. Of particular interest are the positions of the matches to our matrix WM1 representing $\sigma^A$, for these matches define the transcription start sites and thus can be used to determine whether a putative site is bound to by either an activator or a repressor. We defined a regulatory subset *R* of the non-coding sequence to be the regions upstream of the translation starts of divergently transcribed operons. Hence, for each divergent pair of operons, there are two sequences in *R*. We restricted the sequences to be each a maximum of 300 bases, and when the inter-operon distance was < 600 bases, we included the middle overlapping segment (with the appropriate orientation) in both sequences. Because the $\sigma^A$ matrix is far from reverse complement symmetric, this ostensible double counting is not a problem. For comparison, we defined an analogous non-regulatory subset *NR* of the non-coding sequence to be the sequences downstream of the translation stop codons of pairs of convergently transcribed operons. Although the numbers of segements in the two sets are neary equal, there are 175500 independent ways of placing a WM1 match in *R* versus 49500 in *NR*. Hence, the regions between divergent operons are longer than between convergent operons. Still, there are 806 matches to WM1 in *R* versus 99 in *NR*. Hence, the density of matches is 2.2 greater in the regulatory set.

In addition to the greater number of matches, the actual distribution of matches in the regulatory set deviated more from random (see Figure 3). For each set, we defined the random distribution as that expected if each position for a WM1 match in the sequences was equally likely. We then normalized the distribution so that the total number of matches was equal to that observed and binned the counts to obtain the histogram in Figure 3. The deviation between the actual and random distributions was scored with the Kolmogorov Smirnov test [23]. For the interval [-70, -40), there is a 5x greater probability of occurrence of a WM1 match in *R* versus *NR* and 6x greater than for coding sequence (after accounting for the different number of samples). We also looked at all analogous sequences between tandemly transcribed operons, comparing the conventional upstream regulatory region of the downstream operon (*R*) with the same size region immediately downstream of the upstream operon (*NR*). We scored WM1 over the latter region in the opposite sense to transcription to distinguish from perhaps distant sites regulating the downstream operon. Once again, the WM1 matches exhibited a clear positional bias for matching segments in the regulatory set, in particular the [-70, -40) interval. However, the difference between the two sets was less substantial: there was only a 3x greater probability of occurrence of a WM1 match in *R* versus *NR* in the interval [-70, -40).

We tested the matches to the remainder of our 317 matrices for positional biases. For each matrix, we compared the distribution of the matrix's matches in our search space of upstream regions defined in Algorithm with a random distribution defined as the distribution expected if each position for a match were equally likely. Eighteen matrices had biased distributions at a significance level $P < 0.01$ as assessed by the Kolmogorov Smirnov test. The six most significant distributions discounting WM1 are shown in Figure 4. The matches for one of the 18 matrices,

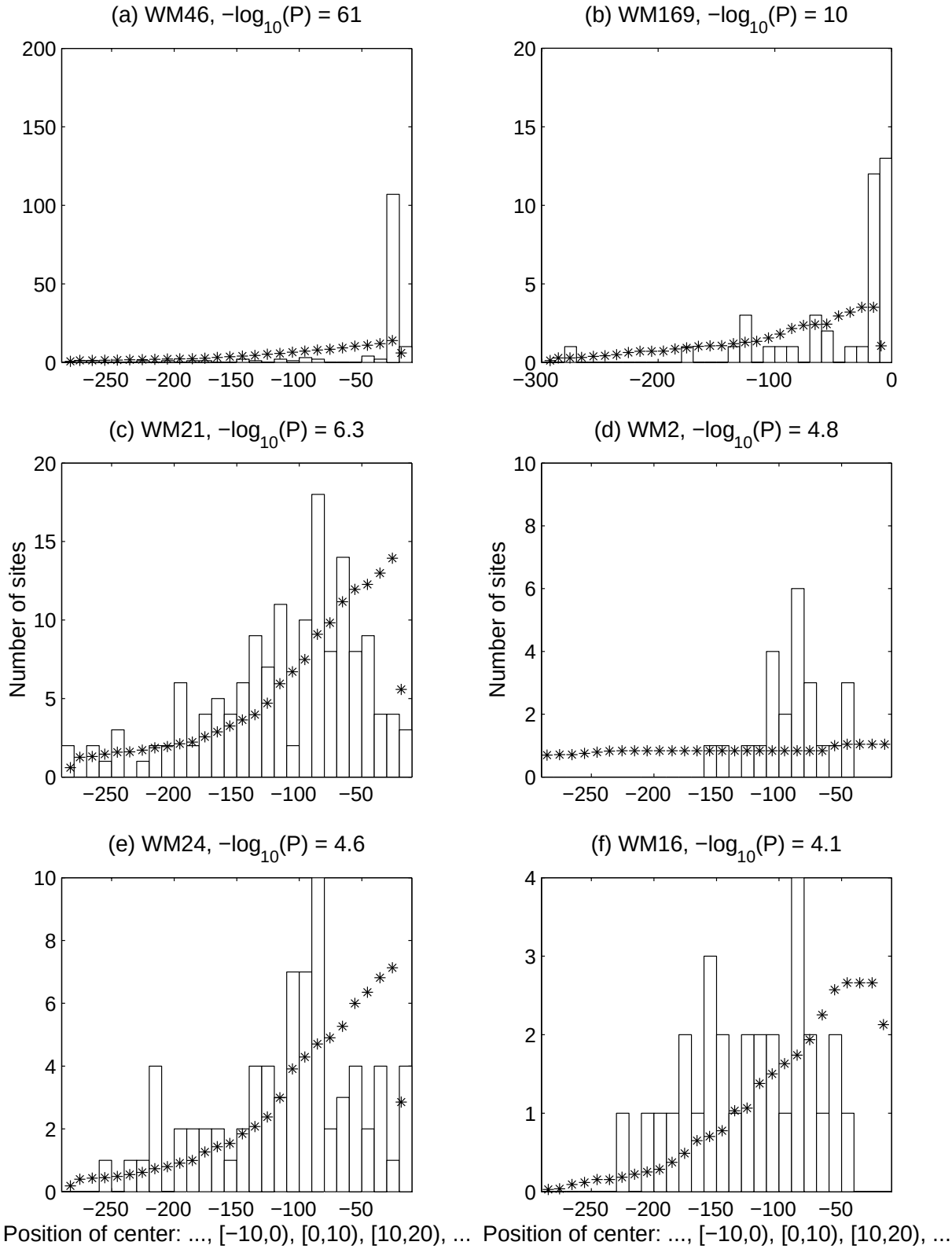


**Figure 4**
Other weight matrices with matches exhibiting a clear positional bias. Histograms of positions of the matches in all upstream sequences to the six non-$\sigma^A$ weight matrices with the most positionally biased matches, using the same conventions as Figure 3.

WM14, tend to occur in the same upstream sequences as the T-box. The c/g richness and the reverse complement symmetry of WM14's consensus -3-CGGC-11-GCCG-3 suggest that the motif is capable of forming a stem loop structure that may interact with the alternate structures formed by the T-box.

A more ambitious test, since it relies on the quality of our $\sigma^A$ predictions, distinguishes matrices representing activators and repressors by their matches' positions relative to the $\sigma^A$ predictions. The position of a matrix match relative to a WM1 match in the same upstream region in our search space is measured center to center. The position -1 indicates that the center of the matrix match is 1 base upstream of the center of the WM1 match. Relative to the center of the $\sigma^{70}$ site in E. coli, the centers of activator sites are concentrated upstream in the -20 to -70 interval, and the centers of repressor sites tend to fall downstream of the -20 position [17]. At a significance level $P < 0.01$, the matches of 13 matrices (excluding those representing $\sigma^A$) exhibited biases relative to our WM1 matches (see Table 2). The six most significant cases are shown in Figure 5, and only WM46 appears to be a repressor. In the case of WM46, the positional bias may come simply from the similarity of the 5' end of WM46, TATA, with the 3' end of WM1 with consensus $N_7TTGAN_{19}TATAATAN_6$. Nevertheless, if WM46 represents an actual factor, it would act as a repressor.

The matches to WM22 did not exhibit a positional bias with respect to our WM1 matches, even though WM22 is a good representative of the canonical repressor LexA (of the 35 matches to WM22, 18 agreed with one of the 30 experimentally verified LexA sites in [36]). In our set of upstream sequences, only 9 of the 35 WM22 matches have a WM1 match to compare with in the same sequence, suggesting that the $\sigma^A$ recognition sites are weak for LexA regulated genes. When a comparison can be made, the centers of the WM22 matches tend to fall upstream of the WM1 matches, which is consistent with the observation that LexA sometimes prevents transcription initiation by binding upstream of the RNA polymerase binding element to inhibit the interaction of the RNA polymerase α-subunit with the a/t rich UP element [36]. A histogram of the positions, again measured center to center, of the 30 experimentally verified LexA sites relative to the known sigma site has the most weight in the upstream interval [-40,-20).

We also checked the weight matrices derived from the experimentally verified recognition sites for the 23 non-sigma factors in the DBTBS database. No matrix had matches exhibiting a positional bias with respect to our WM1 matches at a significance level $P < 0.01$. In a number of cases (e.g. AraR, RocR, CtsR, and HrcA), the total number of matches is small and thus does not define a significant distribution; in other cases (e.g. CcpA, DegU, SpoOA, and TnrA), the regulators act as both activators and repressors; finally, for the factors with more than 400 matches in Figure 2(b), we expect that many of the matches are false positives for reasons stated above.

### *Other applications*

Pathogenicity islands (PAI's) [50] transferred between bacteria present interesting cases for study, for it is not clear if and how the PAI's are coordinately regulated with the host genome. Any cross regulation that exists may not have any profound significance but could occur by chance and not be deleterious. A well studied case is the PAI SPI-1 in S. typhimurium LT2 [51], which encodes two transcription factors HilA and InvF [52] that regulate genes within the island. When we applied our algorithm to the entire S. typhimurium LT2 genome, we did find a marginal match to the HilA recognition consensus but the statistics were poor. There were numerous matrices though that recognized sites within and outside the island, suggesting that the pathogenicity genes are coregulated with the remainder of the genome. We also ran our algorithm on just the SPI-1 island itself but found nothing over-represented that matched the known HilA and InvF recognition consensuses.

## Discussion

There are a number of motif finding algorithms (Consensus [4], Gibbs [5], MEME [6]) that construct a weight matrix directly and are suitable for locating similar patterns in groups of tens of operons. They are thus the best tools for which to process gene clusters obtained from microarrays. (For bacterial applications, their sensitivity is much improved if they fit to dimer patterns with symmetry.) They evaluate significance by reference to a model of random bases (which is far from the truth, even if poly A/T sequences are excluded) and may not converge to the optimal pattern. They also do not use information from beyond the genes being analyzed. Reference [7] search for over-represented monomers of length 6 in a target set. Significance is assessed by contrasting the counts in the target set to the counts genome wide. They are then faced with an assembly problem for the various 6-mers that scored significant and the possibility that a degenerate pattern is significant even when none of the words that overlap it are.

There are a number of algorithms that exploit the dimer symmetry of bacterial motifs [8,10,9]. They differ in how they assign significance. Reference [8] searches for dimers of world length 3 in a subset of sequences and assesses the frequency of occurrence by either contrasting the subset with the genome or using actual word counts and computing a probability from Poisson statistics based on the

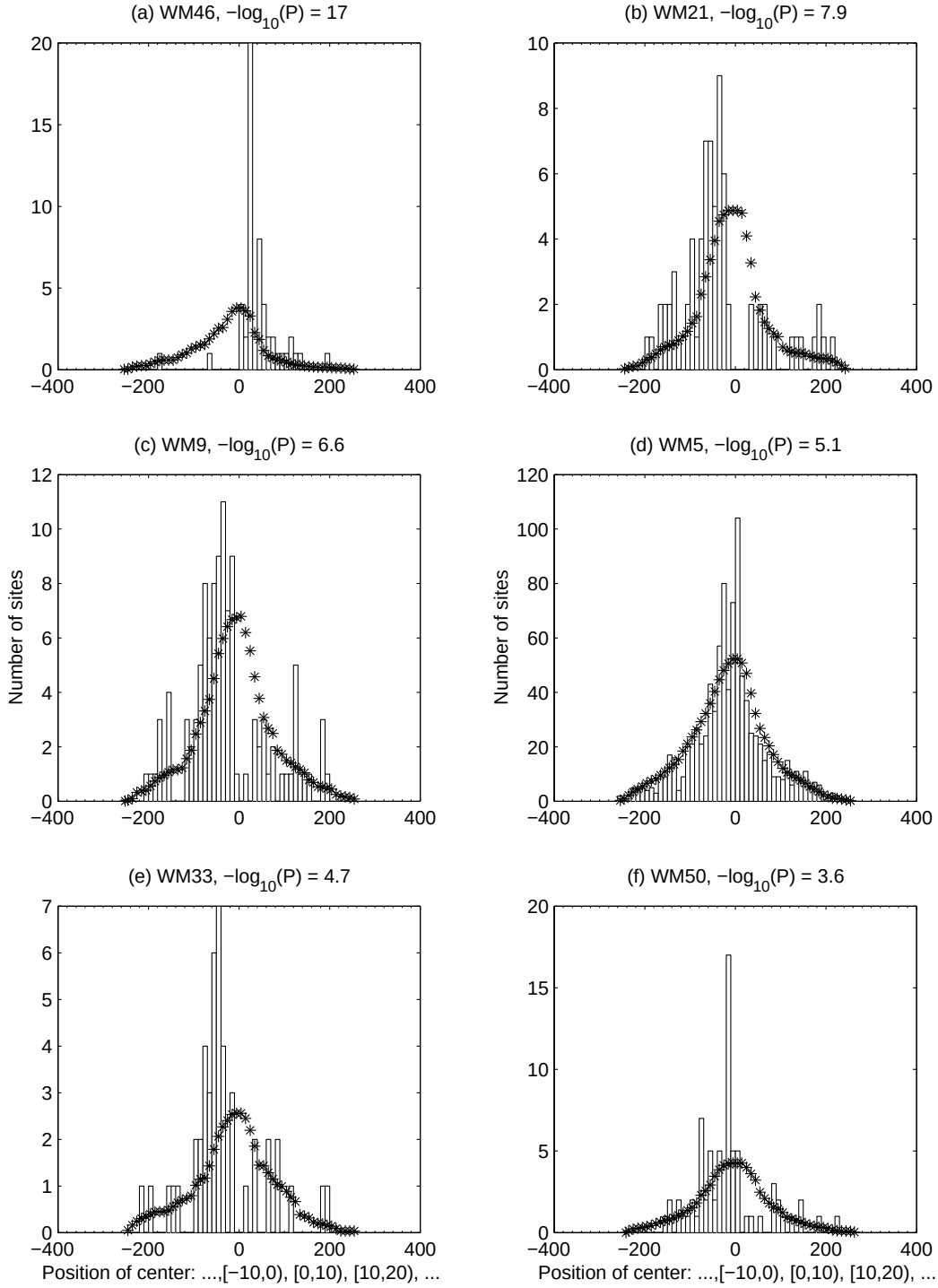


**Figure 5**
Weight matrices with matches exhibiting a clear positional bias relative to $\sigma^A$ sites. Histograms of positions of the matches to weight matrices relative to the best matches to the $\sigma^A$ weight matrix WM1. The position of a weight matrix match relative to a WM1 match is measured center to center. The position -1 indicates that the center of the matrix match is 1 base upstream of the center of the WM1 match. Plots for the six weight matrices with the most positionally biased matches are shown using the same conventions as Figure 3.

spacing, as we do. They do not attempt to cluster the word pairs thereby obtained nor do they attempt genome wide applications. Reference [9] compute significance from a Markov Model applied to the entire dimer. They score degenerate patterns defined by IUPAC symbols and resolve overlapping motifs with a greedy algorithm [53].

Our algorithm is a direct extension of [15], who worked with E. coli. Our principal technical innovations involved the clustering of the dimers, the construction of weight matrices from sites, and the detailed manner in which we validated our predictions using the available B. subtilis information. When applied to E. coli, our clustering procedure gave about half the number of clusters as in the earlier paper (if clusters containing one dimer are counted) and generally reduced the number of nearly equivalent weight matrices. When we computed weight matrices, we did not use low information matrix columns in subsequent scans with the matrix across the genome. This eliminated a certain amount of noise and generally gave us putative regulons of comparable size to the more sophisticated inference method developed by [54].

When we compared our results with the DBTBS and Helmann databases, we hit a smaller fraction (~21%) of the known recognition sites than in the parallel study for E. coli, probably because for many factors only a few sites with a poorly defined consensus are listed. We did do better with sigma sites, and our most significant dimers corresponded to the consensus recognized by $\sigma^A$, the functional homologue of the primary sigma factor $\sigma^{70}$ in E. coli. Although $\sigma^A$ and $\sigma^{70}$ recognize the same consensus, failure to recover the $\sigma^{70}$ matrix in [15] was not due to difference in method but rather the inherent greater variability of the E. coli sites. In contrast to some weight matrix scans in E. coli that generalized from experimental sites, e.g. [19], most of our dimers generated matrices (with the exception of the poly A/T dimers) that gave very reasonable regulon sizes. Our surrogate $\sigma^A$ matrix, WM1, had a regulon of size 1141 and matched 109 out of 132 experimental sites documented by Helmann with the same spacing. Evidence that a matrix represents an actual regulatory factor could be deduced for a total of 52 of our 317 weight matrices using either matches to known sites, correlations in operons functions, or biases in matches' positions. For comparison, the B. subtilis genome is predicted to encode for 200 DNA binding proteins [55]. Some of our predictions may correspond to translation control motifs, which sometimes operate though conserved stems in the mRNA. We tended to set our significance thresholds high so as to minimize false positives. For these reasons, we missed specific factors such as ComA, CtsR, and Fnr. In counting dimers, we insisted that any prediction have a probability < $10^{-7}$ of occurring by chance given our statistical model, i.e. there is about one random prediction among the set of $10^7$ dimers we searched through. A less stringent cutoff could be used for symmetric and reverse complement symmetric dimers since there are fewer cases to examine.

The most serious shortcoming of our algorithm is that it enumerates and scores ACGT words rather than more degenerate patterns. Thus a fairly specific pattern such as the Fur box TGAtAATNATTaTCA, with a variable site in the center of each word, does not yield any single dimer that passes our cutoffs for significance. Another shortcoming is our relatively crude method of predicting operons. A more sophisticated method could be used, like the ones outlined in [56] and [57]. A more fundamental problem is that a transcription factor can distinguish its preferred binding sites in the genome even when it is impossible to discern these sites by searching for statistical over-representation [58]. Interspecies comparisons are an obvious source of additional sequence information, and one can envision a generalization of our counting procedure to handle multiple genomes. Other extensions would use sequence information along with expression or annotation data to assign higher weight to marginal sites falling within a cofunctional gene group.

## Contributions

Both authors contributed to the refinement and implementation of the algorithm and the analysis of the results.

## Acknowledgments

Support for this project was provided by NSF grant DMR 0129848 and by a Natural Sciences and Engineering Research Council (NSERC) of Canada doctoral Julie Payette fellowship to M.M. John Helmann and Tarek Msadek provided guidance on the B. subtilis literature.

## References

1. Sahel B, Bork P and Huang MA: **The identification of functional modules from the genomic association of genes** *Proc Natl Acad Sci USA* 2002, **99:**5890-5895.
2. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO and Eisenberg D: **A combined algorithm for genome-wide prediction of protein function** *Nature* 1999, **402:**83-86.
3. Enright A, Iliopoulos I, Kyrpides NC and Ouzounis CA: **Protein interaction maps for complete genomes based on gene fusion events** *Nature* 1999, **402:**86-90.
4. Hertz GZ and Stormo GD: **Identification of consensus patterns in unaligned DNA and protein sequences: a large-deviation statistical basis for penalizing gaps** *Proceedings of the Third International Conference on Bioinformatics and Genome Research* 1995:201-216.
5. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF and Wootton JC: **Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment** *Science* 1993, **262:**208-214.
6. Bailey TL and Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers** *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology* 1994:28-36.
7. van Helden J, Andre B and Collado-vides J: **Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies** *J Mol Biol* 1998, **281:**827-42.

8. van Helden J, Andre B and Collado-vides J: **Discovering regulatory elements in non-coding sequences by analysis of spaced dyads** *J Mol Biol* 2000, **28:**1808-1818.
9. Sinha S and Tompa M: **A statistical method for finding transcription factor binding sites** *Proceedings of 8th International Conference on Intelligent Systems for Molecular Biology* 2000:344-54.
10. Vanet A, Marsan L, Labigne A and Sagot M: **Inferring regulatory elements from a whole genome. An analysis of Helicobacter pylori sigma(80) family of promoter signals** *J Mol Biol* 2000, **297:**335-353.
11. Gelfand MS., Koonin EV and Mironov AA: **Prediction of transcription regulatory sites in Archaea by a comparative genomic approach** *Nucl Acids Res* 2000, **28:**695-705.
12. Rajewsky N, Socci ND, Zapotocky M and Siggia ED: **The evolution of DNA regulatory regions for proteo-gamma bacteria by interspecies comparisons** *Genome Res* 2002, **12:**298-308.
13. McCue LA, Thompson W, Carmack CS and Lawrence CE: **Factors influencing the identification of transcription factor binding sites by cross-species comparison** *Genome Res* 2002, **12:**1523-32.
14. Blanchette M, Schwikowski B and Tompa M: **An exact algorithm to identify motifs in orthologous sequences from multiple species** *Proc Int Conf Intell Syst Mol Biol* 2000, **8:**37-45.
15. Li H, Rhodius V, Gross C and Siggia ED: **Identification of the Binding Sites of Regulatory Proteins in Bacterial Genomes** *Proc Natl Acad Sci (US)* 2002, **99:**11772-7.
16. Huerta AM, Salgado H, Thieffry D and Collado-Vides J: **RegulonDB: a database on transcriptional regulation in Escherichia coli** *Nucleic Acids Res* 1998, **26:**55-59[http://www.cifn.unam.mx/Computational_Genomics/regulondb/].
17. Gralla JD and Collado-Vides J: **Organization and function of transcription regulatory elements** *In: Escherichia coli and Salmonella: Cellular and Molecular Biology* 2nd edition. Edited by: *Neidhardt FC. Washington, D.C., ASM Press*; 1996:1232-1245.
18. Kurtz S and Schleiermacher C: **REPuter – Fast Computation of Maximal Repeats in Complete Genomes** *Bioinformatics* 1999, **15:**426-7.
19. Robison K, McGuire AM and Church GM: **A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome** *J Mol Bio* 1998, **284:**241-254.
20. Helmann JD and Moran CP: **RNA Polymerase and Sigma Factors** *In: Bacillus subtilis and its Closest Relatives: From Genes to Cells* Edited by: *Sonenshein AL. Washington, D.C., ASM Press*; 2002:289-312.
21. Ben-Dor A, Shamir R and Yakhini Z: **Clustering gene expression patterns** *J Comp Biol* 1999, **6:**281-97.
22. Blatt M, Wiseman S and Domany E: **Super-paramagnetic clustering of data** *Physical Review Letters* 1996, **76:**3251.
23. Press WH, Teukolsky SA, Vetterling WT and Flannery BP: **Numerical Recipes in C** *Cambridge, Cambridge University Press* 1992.
24. Berg OG and von Hippel PH: **Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters** *J Mol Biol* 1987, **193:**723-50.
25. Cavener DR: **Comparison of the consensus sequence flanking translational start sites in Drosophila and vertebrates** *Nucl Acids Res* 1987, **15:**1353-1361.
26. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA and Barrell B: **Artemis: sequence visualisation and annotation** *Bioinformatics* 2000, **16:**944-945.
27. Grundy FJ and Henkin TM: **Synthesis of Serine, Glycine, Cysteine, and Methionine** *In: Bacillus subtilis and its Closest Relatives.-From Genes to Cells* Edited by: *Sonenshein AL. Washington, D.C., ASM Press*; 2002:245-248.
28. Ishii T, Yoshida K, Terai G, Fujita Y and Nakai K: **DBTBS: A database of Bacillus subtilis promoters and transcription factors** *Nucleic Acids Res* 2001, **29:**278-280.
29. Helmann JD: **Compilation and analysis of Bacillus subtilis sigma A-dependent promoter sequences: evidence for extended contact between RNA polymerase and upstream promoter DNA** *Nucleic Acids Res* 1995, **23:**2351-60.
30. Piggot PJ and Losick RL: **Sporulation Genes and Intercompartmental Regulation** *In: Bacillus subtilis and its Closest Relatives: From Genes to Cells* Edited by: *Sonenshein AL. Washington, D.C., ASM Press*; 2002:483-484.
31. Ellinger T, Behnke D, Knaus R, Bujard H and Gralla JD: **Context-dependent effects of upstream A-tracts. Stimulation or inhibition of Escherichia coli promoter function** *J Mol Biol* 1994, **239:**466-75.
32. [http://foodscience.ucdavis.edu/Price_Lab/Bgenes.html].
33. Price CW: **General Stress Response** *In: Bacillus subtilis and its Closest Relatives: From Genes to Cells* Edited by: *Sonenshein AL. Washington, D.C., ASM Press*; 2002:369.
34. Cao M, Kobel PA, Moreshedi MM, Wu MFW, Paddon C and Helmann JD: **Defining the Bacillus subtilis $\sigma^{W}$ Regulon: A Comparative Analysis of Promoter Consensus Search, Run-off Transcription/Macroarray Analysis (ROMA), and Transcriptional Profiling Approaches** *J Mol Biol* 2002, **316:**443-457.
35. Henkin T: **Ribosomes, Protein Synthesis Factors, and tRNA Synthetases** *In: Bacillus subtilis and its Closest Relatives: From Genes to Cells* Edited by: *Sonenshein AL. Washington, D.C., ASM Press*; 2002:319.
36. Dubnau D and Lovett CM: **Tranformation and Recombination** *In: Bacillus subtilis and its Closest Relatives: From Genes to Cells* Edited by: *Sonenshein AL. Washington, D.C., ASM Press*; 2002:463-465.
37. Fisher SH: **Regulation of nitrogen metabolism in Bacillus subtilis: vive la difference!** *Mol Microbiol* 1999, **32:**223-232.
38. Yoshida K, Kobayashi K, Miwa Y, Kang CM, Matsunaga M, Yamaguchi H, Tojo S, Yamamoto M, Nishi R and Ogasawara N: **Combined transcriptome and proteome analysis as a powerful approach to study genes under glucose repression in Bacillus subtilis** *Nucleic Acids Res* 2001, **29:**683-92.
39. Jarmer H, Larsen TS, Krogh A, Saxild HH, Brunak S and Knudsen S: **Sigma A recognition sites in the Bacillus subtilis genome** *Mircobiology* 2001, **147:**2417-2424.
40. Schulz A and Schumann W: **hrcA, the first gene of the Bacillus subtilis dnaK operon encodes a negative regulator of class I heat shock genes** *J Bacteriol* 1996, **178:**1088-93.
41. Baichoo N, Wang T, Ye R and Helmann JD: **Global analysis of the Bacillus subtilis Fur regulon and the iron starvation stimulon** *Molecular Microbiology* 2002, **45:**1613-1629.
42. Britton RA, Eichenberger P, Gonzalez-Pastor JE, Fawcett P, Monson R, Losick R and Grossman AD: **Genome-Wide Analysis of the Stationary-Phase Sigma Factor (Sigma-H) Regulon of Bacillus subtilis** *Journal of Bacteriology* 2002, **184:**4881-4890.
43. Ogura M, Yamaguchi H, Yoshida Ki, Fujita Y, Ogasawara N, Tanaka T and Fujita Y: **DNA microarray analysis of Bacillus subtilis DegU, ComA and PhoP regulons: an approach to comprehensive analysis of B. subtilis two-component regulatory systems** *Nucleic Acids Res* 2001, **29:**3804-13.
44. Kobayashi K, Ogura M, Yamaguchi H, Yoshida K, Ogasawara N, Tanaka T and Fujita Y: **Comprehensive DNA microarray analysis of Bacillus subtilis two-component regulatory systems** *J Bacteriol* 2001, **183:**7365-70.
45. Beier L, Nygaard P, Jarmer H and Saxild H: **Transcription Analysis of the Bacillus subtilis PucR Regulon and Identification of a cis-Acting Sequence Required for PucR-Regulated Expression of Genes Involved in Purine Catabolism** *Journal of Bacteriology* 2002, **184:**3232-3241.
46. Saxild H, Brunstedt K, Nielsen K, Jarmer H and Nygaard P: **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-folD, and pbuO** *Journal of Bacteriology* 2001, **183:**6175-6183.
47. Van Diji JM, Bolhuis A, Tjalsma H, Jongbloed JD, Jong A and Bron S: **Protein Transport Pathways in Bacillus subtilis: a Genome-Based Road Map** *In: Bacillus subtilis and its Closest Relatives: From Genes to Cells* Edited by: *Sonenshein AL. Washington, D.C., ASM Press*; 2002:345-347.
48. Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND and Koonin EV: **The COG database: new developments in phylogenetic classification of proteins from complete genomes** *Nucleic Acids Res* 2001, **9:**22-8.
49. Moszer I, Glaser P and Danchin A: **SubtiList: a relational database for the Bacillus subtilis genome** *Microbiology* 1995, **141:**261-268.
50. Hacker J and Kaper JB: **Pathogenicity Islands and the Evolution of Microbes** *Annu Rev Microbiol* 2000, **54:**641-79.
51. Marcus SL, Brumell JH, Pfeifer CG and Finlay BB: **Salmonella pathogenicity islands: big virulence in small packages** *Microbes and Infection* 2000, **2:**145-156.
52. Eichelberg K and Galan JE: **Differential Regulation of Salmonella typhimurium Type III Secreted Proteins by Pathogenicity**

**Island1 (SPI-1)-Encoded Transcriptional Activators InvF and HilA** *Infection and Immunity* 2000, **67:**4099-4105.
53. Blanchette M and Sinha S: **Separating real motifs from their artifacts** *Bioinformatics* 2001, **17(Suppl 1):**S30-8.
54. Sengupta AM, Djordjievic M and Shraiman BI: **Specificity and robustness in transcription control networks** *Proc Natl Acad Sci USA* 2002, **99:**2072-7.
55. Kunst F, Ogasawara N, Moszer I, Albertini AM, Alloni G, Azevedo V, Bertero MG, Bessieres P, Bolotin A and Borchert S: **The complete genome sequence of gram-positive bacterium Bacillus subtilis** *Nature* 1997, **390:**249-256.
56. Moreno-Hagelsieb G and Collado-Vides J: **A powerful non-homology method for the prediction of operons in prokaryotes** *Bioinformatics* 2002, **18:**S329-36.
57. Ermolaeva MD, White O and Salzberg SL: **Prediction of operons in microbial genomes** *Nucleic Acids Res* 2001, **29:**1216-21.
58. Nimwegen E, Zavolan M, Rajewsky N and Siggia ED: **Probabilistic clustering of sequences: Inferring new bacterial regulons by comparative genomics** *Proc Natl Acad Sci USA* 2002, **99:**7323-8.
59. Tobisch S, Zuhike D, Bernhardt J, Stulke J and Hecker M: **Role of CcpA in regulation of the central pathways of carbon catabolism in Bacillus subtilis** *J Bacterial* 1999, **181:**6996-7004.
60. Fukuoka T, Moriya S, Yoshikawa H and Ogasawara N: **Purification and characterization of an initiation protein for chromosomal replication, DnaA, in Bacillus subtilis** *J Biochem (Tokoyo)* 1990, **107:**732-9.