Motif detection towards the inference of the transcriptional network

Abeer Fadda

Table of contents

CHAPTER 1 ..... **6**

1 INTRODUCTION **6**

1.1 CONTEXT OF THE THESIS 6
1.2 THE TRANSCRIPTIONAL NETWORK 7
1.3 INTEGRATION OF HETEROGENEOUS DATA 9
1.4 SCOPE AND OUTLINE OF THE THESIS 10

CHAPTER 2 ..... **12**

2 MOTIF DETECTION TOOLS **12**

2.1 MOTIF DEFINITION 12
2.2 MOTIF REPRESENTATION 13
2.3 MOTIF FINDING ALGORITHMS 15
2.3.1 Motif finding methods grouped by algorithmic approaches *16*
2.3.2 Motif finding strategies grouped by input data types *21*
2.4 DISCUSSION 27

CHAPTER 3 ..... **29**

3 DELINEATING HILA TARGETS IN *SALMONELLA* TYPHIMURIUM **29**

3.1 INTRODUCTION 29
3.2 MATERIALS AND METHODS 30
3.2.1 ChIP-chip *30*
3.2.2 Transcriptome microarray analysis *31*
3.2.3 Promoter sequences *31*
3.2.4 Running MDscan and MotifRegressor *32*
3.3 RESULTS 32
3.3.1 Motif detection: MDscan *33*
3.3.2 MotifRegressor *35*
3.3.3 Discussion *38*

CHAPTER 4 ..... **40**

4 DETECTION OF DE NOVO MOTIFS IN *BACILLUS SUBTILIS* BY PHYLOGENETIC FOOTPRINTING **40**

4.1 INTRODUCTION 40
4.2 MATERIALS AND METHODS 41
4.2.1 Data set *41*
4.2.2 Retrieving conserved blocks (step 1 in Figure 5.1) *41*
4.2.3 Pair-wise comparison of blocks (step 2 in Figure 5.1) *41*

4.2.4 Clustering of Blocks (step 3 in Figure 5.1) 42
4.2.5 Extraction of motifs (step 4 in Figure 5.1) 42
4.2.6 Screening the B. subtilis genome 42
4.2.7 Calculating functional enrichment and overlap with targets of known regulators 42
4.2.8 Associating de novo motifs with potential regulators 43
4.3 RESULTS 43
4.3.1 General strategy 43
4.3.2 Benchmarking against DBTBS 46
4.3.3 Validation by data integration 51
4.4 DISCUSSION 55

CHAPTER 5 56

5 INFERRING THE TRANSCRIPTIONAL NETWORK OF *BACILLUS SUBTILIS* 56

5.1 INTRODUCTION 56
5.2 MATERIALS AND METHODS 57
5.2.1 Constructing a motif compendium 57
5.2.2 Construction of the microarray compendium 57
5.2.3 Inference of transcriptional regulatory modules 58
5.2.4 Benchmarking of the inferred interactions with DBTBS 58
5.2.5 Functional and condition enrichment 59
5.2.6 Prediction of intra-operonic binding sites 59
5.3 RESULTS 59
5.3.1 Network construction 59
5.3.2 Network modules and inferred interactions 62
5.3.3 Mode of regulation 64
5.3.4 Condition-dependency of the transcriptional network 64
5.3.5 Prediction of intra-operonic motif sites 67
5.3.6 Identification of complex regulons 69
5.3.7 Identifying global regulators 69
5.3.8 Properties of the regulator interaction network 71
5.4 DISCUSSION 75

CHAPTER 6 77

6 CONCLUSIONS AND PERSPECTIVES 77

6.1 CONCLUSIONS 77
6.2 PERSPECTIVES 79
6.2.1 Transcriptional networks involve more complexity than currently accounted for 79
6.3 BRIDGING THE GAP BETWEEN BIOLOGISTS AND BIOINFORMATICIANS 80

REFERENCES 82

APPENDIX 1 ..... **91**

APPENDIX 2 ..... **102**

APPENDIX 3 ..... **108**

Abstract

Bacteria are subject to vigorous environmental changes and their adaptation is crucial for their survival. They achieve this adaptation partly via the intricate regulation of the transcription of their genes. The reconstruction of the transcriptional network is therefore necessary for a full understanding of the dynamics of this regulation. The first step in the building of the network is the identification of the individual regulator-gene interactions. In an attempt to do so, we resolved to detect regulator's binding sites in the promoter regions of genes, and moved on to exploit them in network reconstruction, along the common theme of data integration.

In the first part of the study, we carried out *de novo* motif detection in a set of genes that were bound by the *Salmonella* Typhimurium invasion factor HilA, in a ChIP-chip study, and underwent an expression change in an *hilA* mutant compared to wild type. To this end, the enrichment scores and expression values were used to filter out spurious motifs, by using linear regression analysis. We identified a motif that best explains the expression data, which was found to overlap with the previously detected HilA binding site. Nine genes contained this motif, and were thus identified as HilA targets. They included some previously described targets and some novel ones.

In the second part of the study, we used the Gram-positive model organism *Bacillus subtilis* to perform genome-wide *de novo* motif detection based on an approach that combines phylogenetic footprinting with the concept of co-regulation. The motifs identified were integrated within the framework of DISTILLER to infer the condition-dependent transcriptional network of *B. subtilis*. DISTILLER exploits two types of data: motif and expression data. In addition to the *de novo* motifs, we used known motifs derived from the public database DBTBS, to screen the entire genome and build the motif data. The expression data consists of a compendium of microarrays across different platforms. Our results indicate that a considerable part of the *B. subtilis* network is yet undiscovered; we could predict 417 new regulatory interactions for known regulators and 453 interactions for yet uncharacterized regulators. The regulators in our network showed a preference for regulating modules in certain environmental conditions. Also, substantial condition-dependent intra-operonic regulation seems to take place. Global regulators seem to require functional flexibility to attain their roles by acting as both activators and repressors.

Chapter 1

1 Introduction

1.1 Context of the thesis

The diverse physiological and phenotypic changes that a cell undergoes in its lifetime are governed by gene expression. At the initial step of gene expression, transcription is shaped mainly by the interaction between the RNA polymerase, the transcription factors (TFs) and the promoter sequence of a gene. Although transcription is not the sole determinant of gene expression, it is the bottleneck in this complex pathway. Hence, a full understanding of the interplay between TFs and their target sequences would provide the means to interpret and model the responses of the cell to diverse stimuli. And therefore, the reconstruction of the transcriptional network becomes a vital objective.

Traditional molecular biology methods for resolving the transcriptional regulatory program have relied on the analysis of single genes. These methods, although fairly reliable, are tedious and slow. The need for an efficient 'line of production' of information had led to the 'omics' era. Advances in experimental procedures allowed for the study of hundreds of genes and proteins simultaneously. Terms such as proteomics, transcriptomics, metabolomics, etc, became commonplace. With the flood of information created by the new techniques, came the need for an informatics approach to the problem, also known as *in silico* analysis, which is the topic of this thesis.

The inference of the transcriptional network fits into a bigger field called systems biology (an example can be seen in Figure 1.1). This field's objective is to have a global view on all the structural and functional components of the cell. A good reasoning of systems biology can be understood from the following quote: "the pluralism of causes and effects in biological networks is better addressed by observing, through quantitative measures, multiple components simultaneously, and by rigorous data integration with mathematical models" (Sauer, Heinemann & Zamboni, 2007). Thus, systems biology requires the identification of different biological networks, and the meticulous integration of heterogeneous data sources (omics data). In the next sections we will shed some light on those two areas, namely the transcriptional network and data integration.

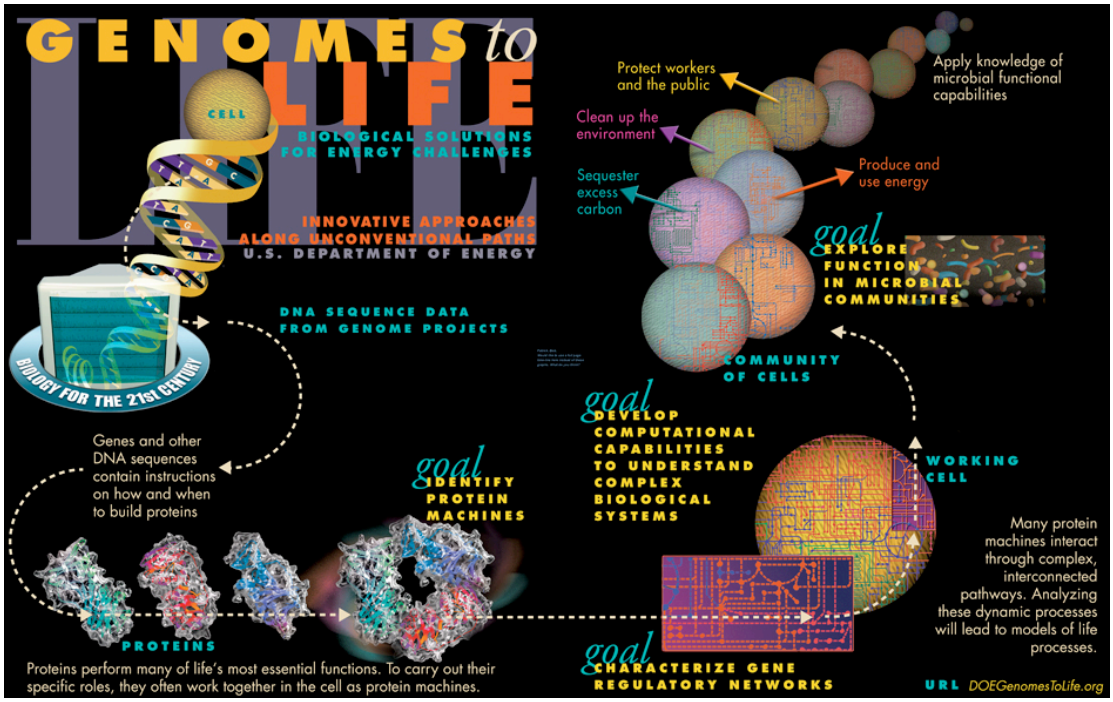


Figure 1.1 An example of systems biology. Systems biology aims at system-level understanding of biological systems, by obtaining, integrating and analyzing complex data from multiple experimental sources. Taken from http://genomics.energy.gov.

1.2 The transcriptional network

The interactions between transcription regulators and their target genes is represented in an abstract way as a graph, in which the nodes signify the regulators and their targets, and the connecting lines, or edges, representing the interaction between them. Because the interactions are non-symmetrical, i.e. a TF regulates a gene but the opposite is not true, the network is said to be directed. In a non-directed network, such as the protein-protein interaction network, there is no direction to the flow of information between nodes. Transcriptional networks are found to be built from smaller units at different levels: starting with the pair-wise interaction between a regulator and its target (basic unit), going through overrepresented simple subgraphs (network motifs), to bigger highly-connected subgraphs (network modules), and ending with the entire network (**Error! Reference source not found.**). Since the basic building unit of the network is the regulator-target interaction, the bulk of this thesis was dedicated to the identification of regulator binding sites (chapters 2 and 3). Note that network motif is different from the sequence motif that is discussed later through the thesis and which refers to the regulator binding site. Network motifs reflect the typical local interconnection pattern in a network, and every real network is characterized with its own motifs (Milo *et al.*, 2002). The characterizing motifs for the transcriptional network

include the single input motif (SIM) describing the connection between a regulator and its target gene, the multiple input motif (MIM) describing the regulation of a gene by several regulators, and the feed forward loops (FFL) describing a situation where the product of a regulated genes acts together with its regulator to regulate the expression of a third gene (**Error! Reference source not found.**). Network modules are a feature of the high clustering of a network. Real networks exhibit a high clustering coefficient and are said to be modular. Generally, the nodes in each module are genes and TFs with similar functions. Thus, delineating network modules can help identifying the functions of uncharacterized proteins. However, in chapter 5, we employ a specific definition of a transcriptional module referring to a set of genes sharing the same motifs and expression profile under a subset of conditions.

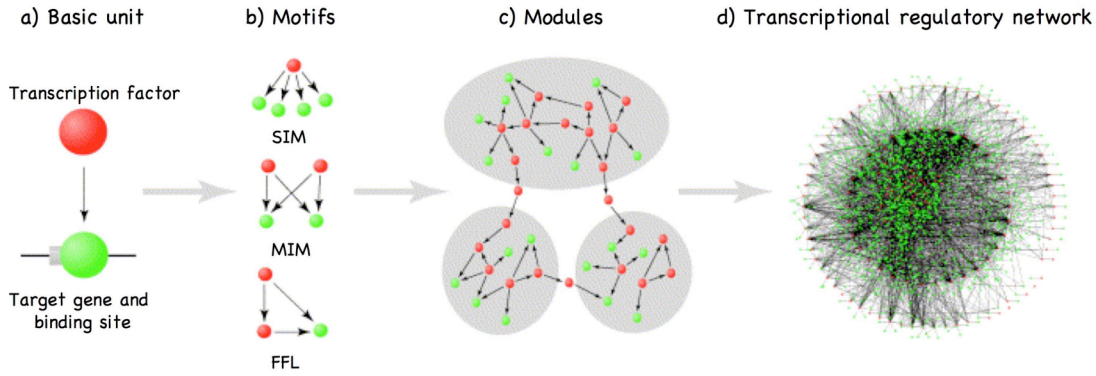


Figure 1.2 Building units of the transcriptional network. a) The basic unit is made of the interacting transcription factor and its target gene. b) Network motifs are overrepresented subgraphs such as single input motifs (SIM), multiple input motifs (MIM), and feed forward loops (FFL). c) Network modules are highly connected subgraphs. d) The entire network is composed of many smaller interconnected units. Taken from Babu *et al.* (2003)

Other features of transcriptional networks include the presence of highly connected nodes called hubs. The topology of the network is such that a few nodes interact with a large number of other nodes, and many nodes interact with just a few. This was found to provide the network with robustness against random attacks, although making is vulnerable to targeted attacks. The presence of hubs also renders the network with the "small world effect", in which the shortest path between two nodes is only a few nodes away. This presumably offers fast communication in the network for a fast and effective regulation.

Transcriptional networks were found to exhibit a hierarchical structure, in which genes are situated in a pyramid-shaped multilayer organization. Through this organization the regulators at the top levels exert a wide effect on the network through regulating other regulators. However, having this structure does not refute the presence of feedback loops through non-transcriptional methods, such as metabolite regulation.

It is important to understand that the topology of the transcriptional network is dependant on environmental conditions. The network was found to undergo rewiring in response to different environmental stimuli (Luscombe *et al.,* 2004); thus, a proper description of the network should include the dynamic aspects of the network.

The networks of *Escherichia coli* and yeast have been reconstructed and their features studied (Babu & Teichmann, 2003; Lemmens et al., 2008; Shen-Orr, Milo, Mangan & Alon, 2002, among others). For this purpose different data sources were used, such as expression profiling, chromatin immunoprecipitation microarrays (ChIP-chip), regulatory-motif data, and comparative genomics. The significance of combining multiple data sources for the reconstruction of the transcriptional network is discussed in the next section.

1.3 Integration of heterogeneous data

As mentioned above, identification of the regulator-gene interaction is the first step in constructing the entire network. Several ways can lead to this end. The discovery of regulator's binding site in a gene can be an indication of a regulatory relationship between the two. Sequence information was utilized by computational methods to identify regulatory binding sites, either on the basis of known regulatory motifs, or by *de novo* detection methods. A drawback to these methods is that they ignore the binding site context and the synergistic or antagonistic contribution of other factors, as well as the similarities of binding sites by related transcription factors. Thus they tend to generate a lot of false positives. Comparative genomics have been used to reduce the noise in this data, based on the observation that functional sequences are more conserved across evolution compared to their non-functional neighborhood. However, sequence conservation should not be over interpreted, as it can be merely a residual of small evolutionary divergence.

A more accurate, though expensive method to identify regulator-gene binding is ChIP-chip analysis. The advantages of using ChIP-chip is that it identifies direct regulator-gene binding *in vivo*, without the need to put the cell through artificial conditions such as regulator overexpression. However, this method has its downsides: false negatives arise because of limited accessibility of the antibody to its epitope and other unknown factors. False positives have also been detected which can be explained by the misleading assumption that the binding of a regulator is sufficient for regulation.

Expression profiling gives information on the end-point effect of a regulator on mRNA levels. The change in the level of a gene's transcript, however, can be due to the change of expression of another regulator whose transcription is controlled by the regulator under study.

Thus, although this method addresses the functionality of a regulator, the affected genes include both direct and indirect targets.

Although the omics data provide a more comprehensive view than reductionist assays, the generation of such massive data comes with a price. Noise constitutes a problem towards an accurate interpretation of the data. Awaiting more refinement of the techniques, the best way to improve the accuracy of interpretation is by combining several data sources, hoping to reduce the false discovery rate. In addition, as discussed above, each of those data reveals a limited view of the network, such that combination of heterogeneous data can reveal a more wholesome picture.

1.4 Scope and outline of the thesis

The objective of this thesis is to identify regulatory binding sites as a first step towards the reconstruction of the regulatory network, supported by an example of network inference in the organism *B. subtilis*. All through the thesis, data integration was put forth. Chapter 2 introduces the subject of *de novo* motif detection, reviewing the many tools developed for this purpose. It starts by explaining motif representation strategies, and moves on to listing of the different motif detection tools developed to date, classified according to algorithmic approaches and input data types. Chapter 3 describes the research done to discover HilA targets in *Salmonella* Typhimurium, using a data integration strategy of ChIP-chip, expression, and sequence data. In chapter 4, phylogenetic footprinting is combined with the concept of co-regulation to detect *de novo* motifs in *B. subtilis*. Those detected motifs are used as part of the input data to the algorithm DISTILLER, which is in turn used to construct the condition-dependent transcriptional network of *B. subtilis* in chapter 5. The data used to infer the network include additionally known regulator binding sites and a compendium of microarrays in diverse conditions. The inferred network is then analyzed and conclusions are made. A schematic of the structure of the thesis can be seen in **Error! Reference source not found.**

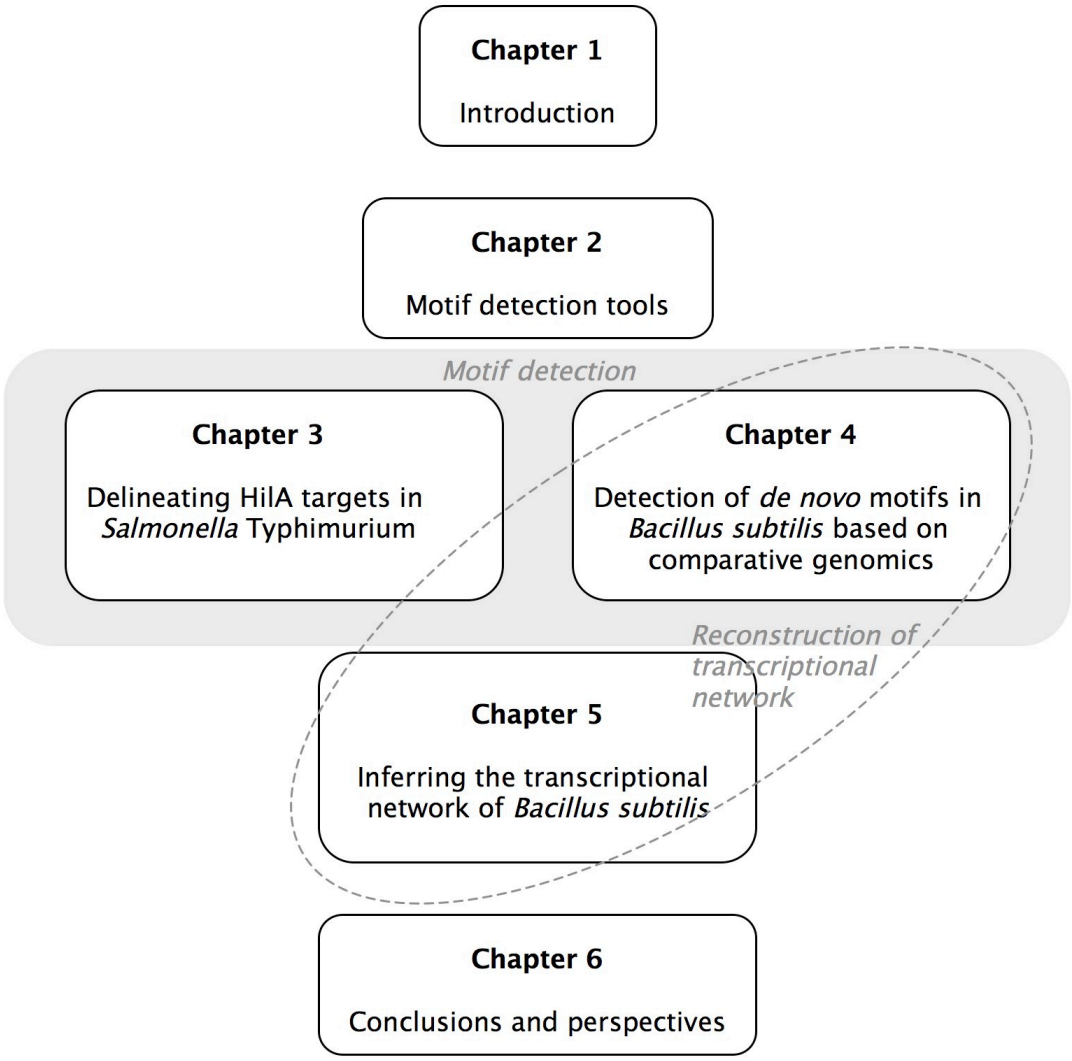


Figure 1.3 Thesis outline

Chapter 2

2 Motif Detection Tools

2.1 Motif definition

This chapter introduces motif representation and reviews the motif detection tools developed to date. As mentioned in chapter 1, gene transcription regulation requires the cooperation of transcription factors with the main transcription machinery. TFs have binding affinities to certain DNA sequences, called motifs or transcription factor binding sites (TFBS). These motifs are generally short (5-20 bp) conserved sequences that recur in different genes, and can be present on either strand of the DNA. In addition to the common single-sequence motifs, spaced dyad motifs also exist, each consisting of two small conserved sites separated by a non-conserved spacer that can be of either a fixed or a variable length. Spaced dyads are an indication that the TF binds to the DNA as a dimer, with two distinct contact points on the DNA. Motifs are generally conserved but can tolerate a degree of degeneracy. The extent of a motif's degeneracy is dependent on the TF; some TFs have less sequence specificity than others. Degeneracy is also a way of fine tuning the transcriptional process, such that the more different a site is from an optimum motif sequence, the less tightly its corresponding gene is regulated by the TF. So, even among the target genes of a TF, we expect to find differences in expression levels due to the differences in their motif sequences. This phenomenon adds a great complexity to the motif detection problem. To start with, a proper way of representing motifs needed to be employed, as will be discussed next.

2.2 Motif representation

A review on motif representation is published by (Stormo, 2000). Four main ways are mostly used: Consensus sequence, Position Frequency Matrix, Position Weight Matrix (or Position Specific Scoring Matrix) and motif logo.

1. Consensus sequence: Each position is shown as one letter representing the most dominant base in that position. For example, the -10 region of the promoter would be represented by the consensus sequence TATAAT. However, it is very rare that this exact sequence is found in promoter regions. A better representation would account for the mismatches or degeneracy of the motif. Thus, the IUPAC (International Union of Pure and Applied Chemistry) nucleic acid codes were employed in which two or more bases occurring at similar frequencies at the same position would be represented by a single letter. Using the same previous example, the -10 promoter region would be represented as TATRNT, allowing for an arginine or a guanine to be present at the 4th position. As much as this representation is an improvement to the 4-letter representation, it is still arbitrary and depends much on convention; for example, a single base is shown if it occurs in > 50% of the sites in some research articles, and in > 60% in others. Yet, this representation is still valid for motif detection tools depending on word enumeration as will be discussed later. The significance of a particular site can be scored given the distribution of all occurrences of the consensus sequence using standard statistical procedures (e.g. Tompa, 1999).

2. Position Frequency Matrix (PFW): In this representation, the frequencies of each of the four DNA bases in the known sites for each of the positions is shown in a matrix. PFMs are more exact representations of the motif and allow for the use of probabilistic methods to search for new sites. However, it assumes a random distribution of the four bases in the genome, which is not the case as genomes are mostly biased in their GC content.

3. Position Weight Matrix (PWM) or Position Specific Scoring Matrix (PSSM): This is a matrix representation of the expected self-information of a particular base in a particular position

$$-f_{b,i} \log f_{b,i} \qquad (2.1)$$

where $f_{b,i}$ is the frequency of base *b* at position *i*.

Pseudocounts have to be added to the frequencies to compensate for the limited observed data and the zero occurrences in the frequency matrix. When the distribution of single bases in the genome are taken into account, the formula becomes as follows

$$-f'_{b,i}\log_2\frac{f'_{b,i}}{p_b} \qquad (2.2)$$

where $f'_{b,i}$ is the frequency of base b at position i with psuedocounts added and p_b is the frequency of base b in the whole genome. Thus, a position's significance (weight) can be measured with this equation

$$I_{seq}(i) = \sum_b f'_{b,i}\log_2\frac{f'_{b,i}}{p_b} \qquad (2.3)$$

which is also a measure of the relative entropy (Kullback-Liebler distance) of the binding site with respect to the background frequencies, and is also equivalent to the log-likelihood ratio. A PWM score of a complete motif is the sum of the log-likelihood scores of all its positions, and thus, it assumes independence between positions of a motif. A PWM is used to search for novel sites with a threshold typically based on the scores of the known sites.

4. Motif logo: This is graphical representation of the motif, where each position is represented by stacks of base letters, the height of which is scaled to the information content (IC) of the base frequency at that position, following this formula:

$$I_i = 2 + \sum_b f_{b,i}\log_2 f_{b,i} \qquad (2.2)$$

where I_i is the information content at position i, $f_{b,i}$ is the frequency of base b at position i. IC indicates how well the base is conserved at each position, and takes a value between 0-2 bits, such that perfectly conserved positions contain 2 bits of information while bases that occur 50% of the time contain one bit. A logo for the -10 region of 60 human promoters, is shown in **Error! Reference source not found.**.

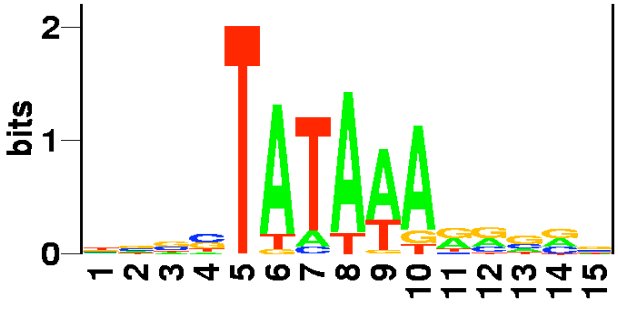


Figure 2.1 Motif logo of the -10 region of 60 human promoters. Taken from http://www.cbs.dtu.dk/staff/dave/roanoke/genetics980320f.htm

Limitations of the mentioned representations:

Two main issues arise with respect to the use of the above motif representations to search for novel sites:

1. Dependence on the number of known sites. The more sites the model is built on, the greater is its accuracy in predicting new sites. This is a major limitation that greatly biases the discovery of new sites, and cannot be overcome except with the laborious biological experiments.

2. Interdependencies of bases within the motif are not accounted for. The significance of this is arguable. While some studies emphasize that interdependencies exist in at least some motifs (Bulyk, Johnson & Church, 2002; O'Flanagan, Paillard, Lavery & Sengupta, 2005), other studies show that accounting for those did not significantly improve the search results (Benos, Bulyk & Stormo, 2002). Several models were suggested to represent interdependencies, e.g. pairwise dependencies (Zhou & Liu, 2004) and Bayesian networks (Barash, Elidan, Friedman & Kaplan, 2003). As complex models maybe better representations of the reality, they come at a cost of needing more data to estimate the parameters, and running the risk of overfitting.

2.3 Motif finding algorithms

Several reviews were published on motif finding algorithms discussing their potentials and limitations (D'haeseleer, 2006; Das & Dai, 2007; GuhaThakurta, 2006; MacIsaac & Fraenkel, 2006; Sandve & Drabløs, 2006). The problem of *de novo* motif finding can be simplified as follows: given a set of genes that we believe to be co-regulated based on evidence such as functional relatedness or co-expression, we assume that they are regulated by the same TF or set of TFs and would like to find their binding sites.

Different classifiers can be used to group *de novo* motif finding strategies: 1) on the basis of algorithmic approaches 2) on the basis of input data types. These classifiers will be used to review the most established motif finding methods.

2.3.1 Motif finding methods grouped by algorithmic approaches

2.3.1.1 Enumerative approaches

Also called word-based methods, enumerative approaches typically do an exhaustive counting of 'words' or short sequences present in the whole search space, and the most frequent words are reported as possible motifs. Consensus motif representation is ideal for these methods. Although first attempts of applying such methods proved to be exhaustive and rigorous (van Helden, André & Collado-Vides, 1998), they were rigid and did not allow for mismatches. Thus, more flexibility was introduced by allowing some fixed number of mismatches (Pavesi, Mauri & Pesole, 2001; Tompa, 1999), or by using the IUPAC codes for 2-nucleotide or 3-nucleotide degenerative positions in the motif (Sinha & Tompa, 2000). However, the computational time for searching the sequence space becomes exponentially bigger with increasing word lengths, such that searching within a search space of 4^L words (L being the motif length) for L > 10 is impractical (Pevzner & Sze, 2000). In general, the computational time complexity is estimated to be $O(NmA^eL^e)$, where N is the number of sequences, m is their length, A is the alphabet size (4 in the case of DNA), L is the motif length, and e is the number of mismatches allowed. Thus, many enumerative approaches had to tradeoff on the number of mismatches to make the searches computationally feasible.

However, some researchers tried to solve this problem by pre-processing the search space to reduce it on statistical basis, making the computation more feasible. Examples of the use of efficient methods for search space reduction include pattern graphs, and projections. In their algorithm WINNOWER, Pevzner *et al*. (Pevzner & Sze, 2000) use graphs to represent all substrings and their similarities, then do a data reduction step by cutting off all spurious edges from the graph. The PROJECTION algorithm reduces the data by using the random projection algorithm (Buhler & Tompa, 2002), splitting the input sequences into 'buckets' such that some bucket receives several occurrences of the desired motif and little else.

Other means for solving the computational efficiency problem for finding long words employed the idea of dictionaries where the 'over-representativeness' of a long word is computed as the weighted average of the short words in the current dictionary which can form partitions of the long word" (Wang, Yu & Zhang, 2005). The first to use the concept of dictionaries in motif search was Bussemaker *et al*. (Bussemaker, Li & Siggia, 2000). In their algorithm MobyDick, they decompose a set of DNA sequences into the most probable dictionary of motifs or words. Then, they identify words by using a probabilistic segmentation model in which the significance of longer words is deduced from the frequency

of shorter ones of various lengths. The model was later extended to encompass motifs or fuzzy words with variable spellings (Sabatti & Lange, 2002; Sabatti, Rohlin, Lange & Liao, 2005). WordSpy (Wang et al., 2005) is a dictionary based algorithm that employs the concept of steganography, where intergenic sequences are seen as a stegoscript in which functional motifs are secret messages embedded in a covertext of background sequences. The algorithm is suitable for genome-wide motif finding, and does not require a prior knowledge of a background sequence model.

Efficient data structures such as suffix trees were also employed in the motif discovery problem. A suffix tree is a data structure that presents the suffixes of a given string in a way that allows for a particularly fast implementation of many important string operations. Any string of length m can be degenerated into m suffixes, and these suffixes can be stored in a suffix tree. Creating this structure requires time $O(m)$ and searching for a motif in it requires time $O(n)$, where n is the length of the motif. The use of suffix trees allows for the search of longer motifs, since the search time is not exponential in the motif length, but in the number of mismatches allowed. Sagot *et al.* (Marsan & Sagot, 2000; Sagot, 1998) were the first to apply suffix trees to search for DNA motifs. Vanet *et al.* (Vanet, Marsan, Labigne & Sagot, 2000) used suffix trees to search for single motifs in whole bacterial genomes. Marsan and Sagot (Marsan & Sagot, 2000) extended the method to search for combinations of motifs, especially applicable to eukaryotic promoters. Eskin and Pevzner (Eskin & Pevzner, 2002) introduced MITRA, an algorithm that borrows the concept of looking for composite motifs from WINNOWER, and uses suffix trees for finding weaker motifs. Weeder (Pavesi, Mereghetti, Mauri & Pesole, 2004) is a web interface for motif search based on suffix trees, which allows for an automatic selection of the values of the parameters needed by the algorithm in a way tailored to the sought motifs, thus making the search more user-friendly. Carvalho *et al.* (Carvalho, Freitas, Oliveira & Sagot, 2006) developed the algorithm RISO, which uses a new data structure, called box-link, to store the information about conserved regions that occur in a well-ordered and regularly spaced manner in the dataset sequences. They claim a time and space gain over the best known exact algorithms.

2.3.1.2 Probabilistic approaches

Such methods use powerful computational techniques to optimize a solution to the motif finding problem. Most of them depend on the PWM representation of the motif. Unlike the enumerative approaches, such methods are not exhaustive, and may get trapped in local maxima.

The first such methods was a greedy probabilistic algorithm developed by Stormo and Hertzell (Stormo & Hartzell, 1989), that was used to search for a motif present once in every sequence, such that for a given set of n sequences, and a motif length L, the algorithm progressively builds matrices by including the sites that maximize the information content. The algorithm was developed later into the CONSENSUS program (Hertz & Stormo, 1999), which includes statistical significance estimation for the IC score.

Another extension of the Hertz greedy algorithm was introduced by Lawrence and Reilly (Lawrence & Reilly, 1990). They used the expectation maximization (EM) algorithm to find protein motifs. The EM algorithm starts with a guess PWM (the expectation step), which can be random or based on some prior knowledge about the binding sites. The probability that each subsequence was generated by the motif and not by the background sequence distribution is calculated, and the PWM is updated based on the weighted average across the probabilities (the maximization step) (Figure 2.2). The algorithm cycles between the two steps until it converges, performing a gradient descent until it reaches the maximum log likelihood of the motif. However, the algorithm may not converge to the global maximum. Thus, a more popular implementation of the algorithm, MEME (Bailey & Elkan, 1995), uses every existing subsequence in the dataset to initialize the iterations, trying to avoid local maxima. MEME also improves on the original EM algorithm by allowing more than one motif occurrence in each sequence.

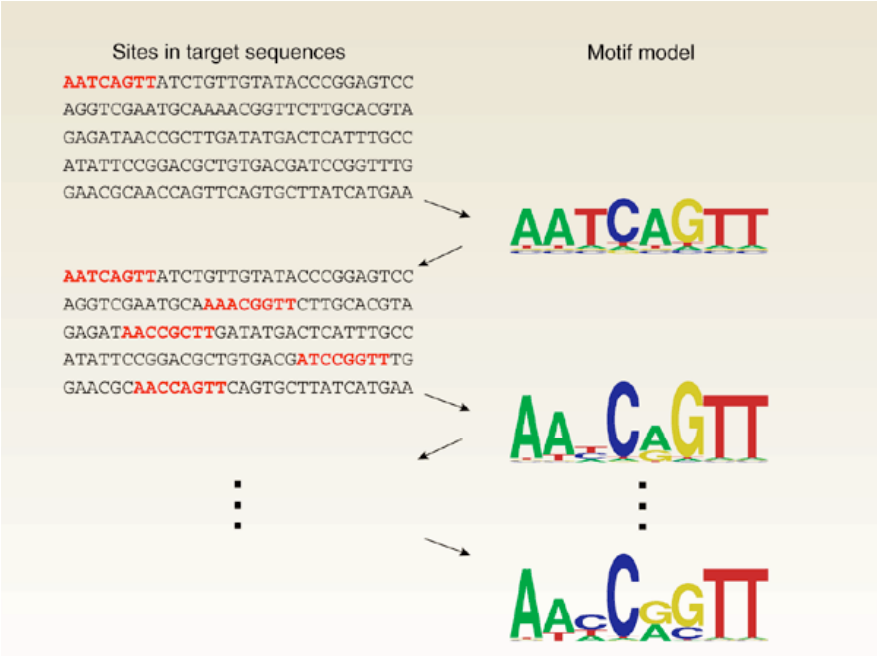


Figure 2.2 Expectation maximization. Starting from a single site, expectation maximization algorithms such as MEME alternate between assigning sites to a motif

(left) and updating the motif model (right). Note that only the best hit per sequence is shown here, although lesser hits in the same sequence can have an effect as well. Taken from D'haeseleer *et al.* (2006).

Another popular probabilistic motif finding algorithm is Gibbs sampling. Gibbs sampling is a Markov Chain Monte Carlo (MCMC) approach, where, similar to a Markov Chain, the results of each step depend only on the result of the previous step, and similar to Monte Carlo methods, each next step is selected by sampling. The first implementation of Gibbs sampling in motif detection was developed by Lawrence *et al.* (Lawrence *et al.*, 1993) to find protein motifs. Later on, the method became widely used in many motif detection software. The motif model is typically initialized with a randomly selected set of sites, and every site is scored against the initial motif model. In the sampling step, a new motif position in one of the sequences is determined by sampling according to a weight distribution of all the sites in that sequence. Since this is a stochastic process, there is no guarantee that the starting position with the highest weight is chosen. The motif model is updated after each sampling step, and the steps are iterated for each of the sequences and until a convergence is reached (Figure 2.3). Many iterations are needed for the algorithm to have efficiently sampled the joint probability distribution of motif models and sites, and reach at the best fitting combinations.

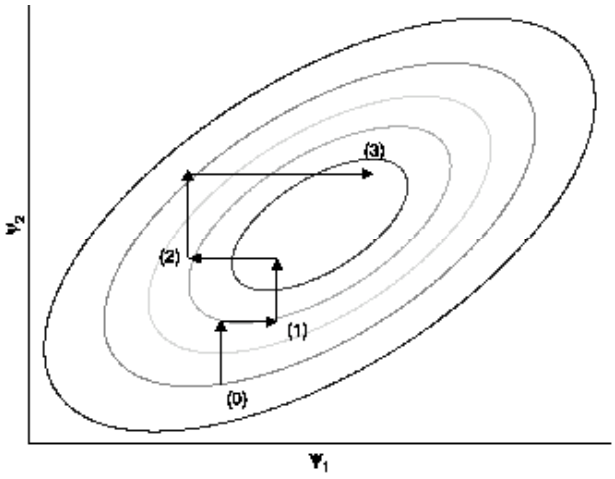


Figure 2.3 Gibbs sampling algorithm in two dimensions starting from an initial point and then completing three iterations. Taken from http://math.u-bourgogne.fr/monge/bibliotheque/ebooks/csa/htmlbook/node28.html

Several algorithms extended the original Gibbs sampling method. The most prominent ones are discussed below.

AlignACE (Roth, Hughes, Estep & Church, 1998) is an algorithm that allows for multiple motifs to be searched by iteratively masking the found motifs. It searches both strands of the DNA simultaneously at each step. Rather than using IC scores and their p-values, AlignACE uses the MAP (maximum *a priori* log likelihood) score, which is the degree of over-representation of a motif compared to its expected random occurrence in the sequence.

ANN-Spec (Workman & Stormo, 2000) utilizes an Artificial Neural Network and a Gibbs sampling method to define the specificity of a DNA-binding protein.

Thijs *et al*. (Thijs et al., 2002) developed MotifSampler, a Gibbs sampling algorithm that uses a probability distribution to estimate the number of copies of the motif in a sequence. It improves on the original Gibbs sampling by taking into account a higher-order Markov-Chain background model, thus increasing the robustness of the algorithm against background noise.

BioProspector (Liu, Brutlag & Liu, 2001) gives the user the liberty to use zero to third-order Markov background model, and the choice between having the model supplied in a separate file or built from the whole genome sequence. It also accounts for gapped and palindromic motifs.

Zhou and Liu (Zhou & Liu, 2004) extended the PWM model to include pairs of correlated positions and design a Markov chain Monte Carlo algorithm called GMS-MP, to sample in the model space.

The problem of Gibbs sampling getting trapped in local optima was tackled by Shida *et al*. (Shida, 2006). Their algorithm, GibbsST, uses the simulated tempering method to improve the Gibbs sampling. Simulated tempering is related to simulated annealing, and is developed in the field of thermodynamics to achieve global optimization.

Recently, Thompson *et al*. presented the Gibbs Centroid Sampler (Thompson, Newberg, Conlan, McCue & Lawrence, 2007). Rather than reporting one optimum solution to the motif finding problem, this method presents a centroid solution, which is an alignment that has the minimum total distance to the set of samples chosen from the *a posterior* probability of a motif's alignment. The centroid solution enhances the sensitivity and positive predictive value of Gibbs sampling.

Some kind of a post-processing of the Gibbs sampling output was suggested by Reddy *et al*. (Reddy, DeLisi & Shakhnovich, 2007). The authors used a graph-based clustering of the aligned instances reported by several hundred runs of the Gibbs sampling with several motif widths. This method takes into account the frequency by which a 'cluster'

of predicted motifs is returned by the Gibbs sampling, where each cluster represents one likely motif.

2.3.1.3 Other methods:

Liu *et al*. and Wei *et al*. use a genetic algorithm (GAs) in their software FMGA and GAME, respectively (Liu, Tsai, Chen, Chen & Shih, 2004; Wei & Jensen, 2006). GAs is a search technique to find exact or approximate solutions to optimization and search problems, inspired by Darwin's theory of evolutionary biology and borrows such features as inheritance, mutation, selection, and crossover. FMGA uses PWM to perform mutations, and rearrangements to avoid local optima. GAME employs a PWM-based Bayesian model, and two operations, ADJUST and SHIFT, to avoid local optima.

Fratkin *et al*. (Fratkin, Naughton, Brutlag & Batzoglou, 2006) present the problem of motif searching in a graph approach. In their software, MotifCut, they drop the PWM representation of the motif and use a graph presentation where the vertices represent all k-mers in the input sequences, and the edges represent the pairwise k-mer similarities. They search for motifs by looking for maximum density sub-graphs. Their method models the inter-motif dependencies of the bases.

Kingsford *et al*. (Kingsford, Zaslavsky & Singh, 2006) use integer linear programming (ILP) to find subsequences of a given length such that the sum of their pairwise distances is minimized.

2.3.2 Motif finding strategies grouped by input data types

Here we review motif-finding methods that strictly require other data types besides promoter sequences.

2.3.2.1 Structural information

Relatively, very few studies used structural data for motif prediction, due two reasons: the complexity of applying a structure-based TF-DNA binding model, and the scarcity of structural data since it is not easily generated in the wet lab.

Mandel-Gutfreund *et al*. and Kaplan *et al*. (Kaplan, Friedman & Margalit, 2005; Mandel-Gutfreund, Baron & Margalit, 2001) used solved protein–DNA complexes to determine the exact architecture of interactions between nucleotides and amino acids at the

DNA-binding domain. They inferred the context-specific amino acid–nucleotide recognition preferences and used them to predict novel binding sites.

Pudimat *et al*. (Pudimat, Schukat-Talamazzini & Backofen, 2004) incorporated helical parameter features in a Bayesian net that is specific for each motif.

Morozov *et al*. (Morozov, Havranek, Baker & Siggia, 2005) developed a physical energy function, which uses electrostatics, solvation, hydrogen bonds and atom-packing terms to model direct readout and sequence-specific DNA conformational energy to model indirect readout of DNA sequence by the bound TF.

Liu *et al*. (Liu, Guo, Li & Xu, 2008) developed a structure-based prediction approach, which integrates a protein-DNA docking algorithm.

2.3.2.2 Co-expression (microarrays)

Bussemaker *et al*. (Bussemaker, Li & Siggia, 2001) used a simple linear regression model in which upstream motifs contribute additively to the log-expression level of a gene. All the genes are simultaneously fit and the statistically significant models that best fit the expression data are selected.

Conlon *et al*. (Conlon, Liu, Lieb & Liu, 2003) developed the algorithm MOTIF REGRESSOR in which they perform a step-wise regression analysis of the scores of *de novo* motifs with their expression level in microarray experiments.

Chen *et al*. (Chen, Hata & Zhang, 2004) predicted motif sites in mutant expression data. They first clustered the differentially expressed genes according to their functional category, then used a word-counting algorithm to detect motifs in each cluster. Their method, allegedly excludes any downstream targets of the knocked out TF from the search space by using the functional clustering.

Dai *et al*. (Dai, He & Zhao, 2007) developed an algorithm that utilizes Support Vector Machines (SVM) to build a classification problem of targets and non-targets. They use a group of genes shown to be co-regulated by a certain TF in a microarray study, as their positive training set, and another group of genes that actually contain the motif but do not belong to the first group, as their negative training set. They use the method to discover new targets of the TF.

MotifScorer (Brilli, Fani & Lió, 2007) is another program that uses regression on motif scores and expression levels. It goes further to predicting condition-specific motifs by relying on a compendium of microarray data across a number of biological conditions.

Chen at al. (Chen, Guo, Fan & Jiang, 2008) suggested the integration of gene expression data into their original algorithm AlignACE, by looking for a PWM such that the likelihood of simultaneously observing both binding sequences and their associated gene expression data is maximized.

A number of methods (Gupta & Liu, 2005; Marsan & Sagot, 2000; Segal *et al.*, 2003; Sinha, Liang & Siggia, 2006) search for cis-regulatory modules (CRM) in co-expressed gene sets, answering to the fact that motifs, especially in higher eukaryotes, tend to occur as homotypic or heteroptypic combinations. The search for such combinations gives better predictions than searching for single motifs.

2.3.2.3 ChIP-chip and universal protein-DNA microarrays

MDscan (Motif Discovery Scan, Liu, Brutlag & Liu, 2002) combines two motif search strategies, word enumeration and position-specific weight matrix updating, and incorporates the ChIP-array ranking information to accelerate and enhance the search for motifs.

Eden *et al.* (Eden, Lipson, Yogev & Yakhini, 2007) developed the software DRIM (Discovery of Rank Imbalanced Motifs), which utilizes ranked lists of sequences, in this case ChIP-chip and CpG methylation data, to search for motifs. The algorithm partitions the data into top and bottom in a way that maximizes the statistical significance of the discovered motif. They employ the minimal hypergeometric score (mHG) for their motifs.

Recently, Berger *et al.* (Berger *et al.*, 2006) developed a protein binding microarray (PBM) that contains synthetic DNA representing all possible sequence variants of k length on a single, universal microarray. The technique was exploited by Chen *et al.* (Chen, Hughes & Morris, 2007) in their algorithm RankMotif++, to predict binding sites of a TF. The algorithm learns motif models by maximizing the likelihood of a set of binding preferences under a probabilistic model of how sequence-binding affinity translates into binding preference observations. It makes use of the entire set of data rather than the highest-ranking data only.

2.3.2.4 Orthologous sequences

The availability of many whole genome sequences inspired comparative genomic studies for motif searching. Other terms used to describe this general category of methods are phylogenetic footprinting and phylogenetic shadowing, although some make a distinction between the two terms. The main idea behind these methods is the assumption that a selective

evolutionary pressure is exerted on sequences, such that, within non-coding sequences, functional elements are preferably conserved relative to the non-functional surroundings. The term phylogenetic shadowing was first introduced by Boffelli *et al*. (Boffelli *et al*., 2003), which, according to their definition, differs from phylogenetic footprinting by the fact that it examines sequences of closely related species and takes into account the phylogenetic relationships of the set of species analyzed. A more detailed explanation of the principles of phylogenetic footprinting is discussed in chapter 4.

Since its first use in 1988 (Tagle *et al*., 1988) until today, several algorithms were developed for phylogenetic footprinting. Besides their differences in the underlying algorithmic basis, the main distinctions between them can be summarized as follows:

1. The use of alignment algorithms versus general motif discovery algorithms. Methods that use alignments make use of the fact that regulatory sites that occur in sets in the promoter region, such as CRM's in eukaryotes, are mostly conserved in their order, orientation and number across orthologous sequences. However, since this is not true in all cases, several methods adapted their alignment strategies. Methods that employ general motif discovery algorithms, like the ones discussed in the previous section, are more suited for prokaryotes, where motif combinations are less frequent.
2. The use of sequence data alone, versus the additional use of co-expression data. Sequence-alone methods have the advantage of not needing additional data sources, which may not be readily available. Sequence-alone methods can be applied for a single gene as long as its orthologs are available. On the other hand, methods that combine co-expression data provide higher accuracy.
3. The use of phylogenetic distances. Most recently published methods include phylogenetic distances. The necessity of including this info comes from the use of multiple (more than two) species comparisons where the phylogenetic distances between them are not uniform. This would lead to biased results towards motifs in the more closely related species.

Here is a quick review of some of the phylogenetic tools in a chronological order.

The first use of comparative genomics to delineate TF binding sites was done by Tagle *et al* in 1988 (Tagle *et al*., 1988), where they were able to predict evolutionary conserved motifs responsible for embryonic ε and γ globulin gene expression in primates.

Gelfand *et al*. (Gelfand, Koonin & Mironov, 2000) and McGuire *et al*. (McGuire, Hughes & Church, 2000) used a mixed set of co-regulated and orthologous genes to search for motifs.

Loots *et al.* (Loots *et al.*, 2000) applied phylogenetic footprinting to discover a coordinate regulator of interleukin genes in human and mouse sequences.

Blanchette *et al.* (Blanchette, Schwikowski & Tompa, 2000) included the phylogenetic distances between the organisms for the first time. They presented the problem as a Substring Parsimony Problem. They applied their method to find motifs across a family of 10 plants and 12 Drosophila species independently.

McCue *et al.* (McCue *et al.*, 2001) applied whole genome footprinting on six gamma proteobacterial genomes using a Gibbs sampling method.

Blanchette and Tompa (Blanchette & Tompa, 2003) extended their original algorithm to include phylogenetic trees in the new method FootPrinter.

Kellis *et al.* (Kellis, Patterson, Endrizzi, Birren & Lander, 2003) looked for motifs in a sequential manner: first they found conserved motifs, secondly they looked for the over-represented ones among them.

Wang and Stormo (Wang & Stormo, 2003) developed the algorithm PhyloCon which also works in two steps: they first look for conserved regions among orthologous sequences, then they compare the found 'profiles' within one species and gradually merge models of similar profiles, making use of the fact that within-species co-regulated genes share the same motifs. Their method was later adapted so that it would first align the found motif models then cluster them (PhyloNet; Wang & Stormo, 2005).

Cliften *et al.* (Cliften *et al.*, 2003) did a whole genome study on six yeast genomes, using ClustalW for alignment.

CONREAL (CONserved Regulatory Elements anchored Alignment) is an algorithm developed by Berezikov *et al.* (Berezikov, Guryev & Cuppen, 2005; Berezikov, Guryev, Plasterk & Cuppen, 2004). The assumptions behind CONREAL are that the sequence and order of functional regulatory elements are mostly conserved in orthologous promoters. It uses known PWMs to screen potential transcription-factor binding sites, and uses them to establish anchors between orthologous sequences and to guide promoter sequence alignment.

Prakash *et al.* (PRAKASH, BLANCHETTE, SINHA & TOMPA, 2003) developed OrthoMEME. The algorithm looks for motifs in combined sets of co-regulated and orthologous genes. This method is different from that of Wang and Stormo in that it looks in the two sets simultaneously rather than sequentially.

EMnEM (Expectation-Maximization on Evolutionary Mixtures, MOSES, CHIANG & EISEN, 2003) is a method that uses an evolutionary model to extract motifs from a mixed set of co-regulated and orthologous gene sets.

PhyMe (Sinha, Blanchette & Tompa, 2004) also uses a mixed set of sequences with the advantage of having the flexibility of leaving out conservation in cases where the motif is not conserved, and finding motifs in co-expressed genes only.

Elnitski *et al.* and Kolbe *et al.* (Elnitski *et al.*, 2003; Kolbe *et al.*, 2004) developed tools to distinguish regulatory sequences from 'neutral' ones across human-rodent sequence alignments. They distinguish 'the regulatory potential' of sequences based on the pattern of observed identical nucleotides. For example, unlike non-coding regions, coding sequences tend to have degeneracy at the third codon position and the insertions and deletions tend to be in multiples of three.

PhyloGibbs (Siddharthan, Siggia & van Nimwegen, 2005) searches for motifs in a mixed set of co-regulated and orthologous sequences simultaneously, taking phylogenetic distances into consideration.

Jensen *et al.* (Jensen, Shen & Liu, 2005) use a two-step approach similar to that of Wang and Stormo (Wang & Stormo, 2003), where conserved motifs are discovered first, then clustered assuming co-regulation of the corresponding genes in each cluster.

BlockSampler (Monsieurs *et al.*, 2006) is an adaptation of MotifSampler for phylogenetic footprinting. The algorithm was modified so that the motif length is not user-defined. Once a seed motif is found, it is extended on both sides of the sequence until the consensus score drops below a set threshold. The user has to specify one of the input sequences as a reference sequence to do the footprint; a motif must exist in the reference sequence and in at least one other sequence, but not necessarily in all the compared sequences. Another adaptation of the algorithm was the use of a species-specific background model for each sequence, since each sequence comes from a different species.

Sosinsky *et al.* (Sosinsky, Honig, Mann & Califano, 2007) developed the tool EDGI (Enhance Detection using only Genomic Information). The algorithm allows the identification of large regulatory elements, such as enhancers, as evolutionary-conserved order-independent clusters of short conserved sequences. They do not model phylogenetic distances. Contrary to global alignment methods, they allow for detection of CRM's that were not conserved in their order in the promoter, or their orientation, or even underwent duplication or deletion of some elements in some species.

Similar to the Gibbs Centroid Sampler (Thompson et al., 2007), a phylogenetic Gibbs sampler was introduced (Newberg *et al*., 2007) which presents centroid solutions for the motif finding problem across orthologous sequences, taking an evolutionary model into account.

WeederH (Pavesi, Zambelli & Pesole, 2007) adds a positional conservation constraint on the motif discovery problem, so that motifs should be conserved in their position in the orthologous genes. They also address the problem in such a way that takes into account the different evolutionary rates for different genes; they look for motifs that are significantly different in their conservation than the surrounding.

2.4 Discussion

The problem of motif finding remains challenging. Most of the methods present today suffer from high rates of false positives and low sensitivity (Wasserman & Sandelin, 2004). This is due in part to our incomplete understanding of the biology of the transcription process. However, some of the assumptions underlying these methods are inaccurate and would lead to faulty results. For example, most of the available methods try to find over-represented motifs; in the long eukaryotic promoter regions the motifs are very subtle signals and over-representation does not hold. Additionally, many methods wrongfully assume that binding of a TF to its target sequence is independent from the binding of other proteins or from the DNA context, leading to the discovery of non-functional binding site. Another wrong assumption on which some methods that rely on expression data use, is that co-expression is a result of transcriptional co-regulation. Co-expressed genes can either be direct or indirect targets of the regulator, thus co-expression does not guarantee the presence of a binding site in all of the genes. On the other hand, the measured levels of RNA are a result of several cellular control mechanisms besides transcriptional initiation, such as transcriptional termination, and mRNA degradation; this may lead to actual targets of the TF being missed because of lower mRNA levels.

Nonetheless, some studies have shown that combining several motif search algorithms and choosing the most prominent of their motif prediction results would increase the specificity and sensitivity of motif detection (Hu, Li & Kihara, 2005; Tompa et al., 2005). For example, the use of phylogenetic footprinting as a first step would significantly decrease the search space for a subsequent motif search (Cliften et al., 2003). This is the strategy that we employed in chapter 4 of this thesis to find *de novo* motifs in *B. subtilis*.

Chapter 3

3 Delineating HilA targets in *Salmonella* Typhimurium

3.1 Introduction

Several methods for motif detection have been described in chapter 2. The importance of *in silico* motif discovery in the biological world will be demonstrated in this chapter by applying one of those methods for the detection of motifs in Salmonella. Here, a motif was detected for one of the key regulators (HilA) of the invasion pathway in the species *Salmonella enterica* serovar Typhimurium. The work was performed in collaboration with Dr. Inge Thijs in the CMPG department and was part of the work published in Journal of Bacteriology (Thijs *et al*., 2007).

Salmonella is a gram-positive facultative anaerobic bacterium, belonging to the family Enterobacteriaceae. It is a rod-shaped bacterium, mostly with peritrichous flagella. The genus is divided over 2,000 serotypes based on their surface antigens. Of the infamous salmonellae, S*almonella enterica* serovar Typhi is the causative agent of typhoid fever, which has come under control in the modern world in the last century. Other serotypes, however, remained problematic, such as *Salmonella enterica* serovar Typhimurium and S*almonella enterica* serovar Enteritidis, the most common causes of salmonellosis. Salmonellosis is a gastrointestinal infection ranging in severity from a self-contained infection to a life-threatening disease, depending on the host's physical condition and the number of pathogenic cells ingested. In humans, symptoms include diarrhea, fever, vomiting, and abdominal cramps 12 to 72 hours after infection and may last for up to 7 days. The method of infection is by oral intake of contaminated foods, mainly milk, eggs and meats.

Inside the intestines, bacterial response to the change in the environment triggers the secretion of virulent factors, which in turn induce changes in the epithelial lining of the intestines, causing internalization of the bacterial cells (or invasion). Once internalized, bacteria survive and replicate inside the Salmonella-containing vacuoles (Ramsden, Mota, Munter, Shorte & Holden, 2007).

A number of bacterial agents mediate the invasion or induced phagocytic process. These agents are practically injected into the host cytosol via a needle-like structure, called type-III secretion system (TTSS). Expression of TSS genes is kept under tight control by HilA among other actors. HilA is a member of the OmpR/ToxR family of transcriptional regulators. Some of the known direct targets of HilA regulation are its own gene, the *inv*/*spa* and *prg* operons, encoding components of the TTSS complex, and the *sic*/*sip* operon, encoding a chaperon and secreted proteins. In this work, a genome-wide approach is used to delineate direct targets of HilA. Data from chromatin immunoprecipitation microarray (ChIP-chip) experiments are combined with those of an expression microarray of a *hilA* mutant, and motif detection output to delineate the direct targets of HilA.

3.2 Materials and methods

3.2.1 ChIP-chip

ChIP-chip was performed for three biological replicates of both the SL1344 wild-type strain and the CMPG5805 HilA-tagged strain (Thijs et al., 2007). Salmonella strains were cultured under high-osmolarity and limited-aeration conditions until an optical density at 595 nm of 0.4 (1E-09 CFU/ml) was reached. Chromatin immunoprecipitation was performed essentially as described by Laub et al. (Laub, Chen, Shapiro & McAdams, 2002) and Shin and Groisman (Shin & Groisman, 2005).

For the array analysis, immunoprecipitated DNA fragments were blunted and amplified via ligation-mediated PCR (LM-PCR) (Mueller & Wold, 1989). For the macroarray analysis, DNA fragments were labeled with digoxigenin-dUTP in LM-PCR, and hybridized to a macroarray containing the intergenic regions of known and hypothesized HilA target genes, i.e., *prgH, invF, sicA, hilA*, and *siiA*, along with a number of negative controls. For the microarray analysis, LM-PCR-amplified DNA fragments were indirectly labeled with Cy5 as described in Oberley *et al.* (Oberley, Tsao, Yau & Farnham, 2004). DNA from cells undergoing the same protocol without the immunoprecipitation step was labeled with Cy3 and used as a common reference for all hybridizations. The applied Salmonella microarrays

covered 99.4% of the *S. enterica* serovar Typhimurium LT2 genome non-redundantly supplemented with genes from strains *S. enterica* serovar Typhi CT18, *S. enterica* serovar Paratyphi A SARB42, and *S. enterica* serovar Enteritidis PT4; each array element, representing a coding sequence, was spotted in triplicate. Data were linearly rescaled using an orthogonal regression fit. To avoid increasing variation in the low-intensity range, no background correction was performed. To detect significantly enriched genes, a two-sample t-test was applied to compare log ratios for the ChIP-chip samples from the CMPG5805 and wild-type strains.

3.2.2 Transcriptome microarray analysis

Salmonella strains SL1344 and CMPG5804 were cultured under high-osmolarity and limited-aeration conditions to induce *hilA* expression (Rodriguez, Schechter & Lee, 2002) until an optical density at 595 nm of 0.4 was reached. Total RNA was isolated and contaminating genomic DNA was removed from the RNA samples. Prior to labeling, the concentration of total RNA was determined by measuring the absorbance at 260 nm. RNA was labeled with Cy5 and Cy3 by reverse transcription (Wang, Frye, McClelland & Harshey, 2004). Hybridizations were performed using a color flip technique and *S. enterica* serovar Typhimurium arrays containing 70mer oligonucleotides representing all LT2 annotated genes (Operon) spotted in duplicate on CodeLink activated slides (Amersham Biosciences). Data were Loess normalized, and no background correction was performed. Differentially expressed genes were detected by significance analysis of microarrays (SAM) (Tusher, Tibshirani & Chu, 2001). A d-value representing a measure of differential expression (d_i) was calculated as $r_i/(s_i + s_0)$, where i is 1, 2, ... , p genes. The parameter s_i represents the standard deviation, s_0 a fixed factor, and r_i a score which is the mean of the log ratios in the one-class case we applied.

3.2.3 Promoter sequences

Salmonella typhimurium LT2 GeneBank (NC_003197) file was used to obtain promoter sequences of the corresponding STM genes. For SL2674 gene (*sopE*) the promoter sequence was taken from Sanger institute *S. enterica* serovar Typhimurium SL1344 annotation (http://www.sanger.ac.uk/Projects/Salmonella/SL1344_web.tab). Promoter sequences correspond to non-coding intergenic regions, except for *sopE* where an additional 100 bp were included on both sides.

3.2.4 Running MDscan and MotifRegressor

The package was downloaded form the website http://www.math.umass.edu/~conlon/mr.html. The regression program "regressExpr.spl" was translated into the R programming language to fit into our working environment.

A list of 19 genes comprising the intersection of positive hits in the ChIP-chip experiment and in the microarray experiment were used. The list was ranked according to the p-value of the ChIP-chip experiment, and this ranking was used for motif detection by MDscan. The 10 most enriched (lower p-values) genes were used for the seed motif detection, and the entire list was used for refinement. The microarray d-values of the SAM test were used as the response factor in the regression analysis (MotifRegressor). No background sequence distribution file was provided and thus the background distribution was calculated by the algorithm based on the input intergenic sequences. Other parameters used were as follows: maximum motif width = 20; minimum motif width = 5; seed candidate motifs to find = 10; motifs to report before regression = 5.

3.3 Results

Identifying targets of a TF on a genome-wide scale is not a trivial task. Two main techniques, microarrays and CHIP-chip are often used for that purpose. The two techniques can be seen as measuring the TF regulation process at different stages; while ChIP-chip measures the initial step of binding of the TF to its target sequence, a microarray measures the outcome of that binding on mRNA production. The two measurements would be redundant if the two events have a strict one-to-one relationship, i.e. if each binding event leads to regulation of expression and if regulation of expression takes place only upon binding. But the reality is different. TF binding to a sequence does not always result in regulation of expression due to factors such as co-regulation by other TFs or negative regulation by sRNAs for instance. Similarly, regulation of expression does not strictly result from TF binding to a gene since a gene can be a secondary target of the TF. Thus, the two techniques can be used complementarily to delineate the direct targets of a TF on a genome level.

To identify the targets of the *S. enterica* serovar Typhimurium HilA regulator, a ChIP-chip experiment was complemented with transcriptome analysis (microarray) of a *hilA* mutant versus the wild type. The ChIP-chip experiment was performed in HilA inducing conditions (see materials and methods) in three biological replicates for both a HilA-tagged strain and a

non-tagged strain. Sequence enrichment for HilA binding sites was expressed in terms of p-values, with positive hits having a p-value < 0.05. Expression profiling of a wild-type strain versus an *hilA* mutant strain was performed under the same conditions as the ChIP-chip analysis. Differentially expressed genes were identified using the statistical test SAM (Significance Analysis of Microarrays), which assigns a score (d-value) for each gene on the basis of the change in its expression relative to the standard deviation of repeated measurements for that gene (Tusher et al., 2001). An absolute d-value > 2 was considered significant. The details of both assays can be found in Thijs *et al*. (2007). For further validation of the genome-wide approach, the *in vitro* binding of HilA to a selection of these putative direct targets was assessed with an electrophoretic mobility shift assay (EMSA). This confirmed the binding of HilA to the previously known targets *prgH, hilA, sicA* and *invF*, and the new targets *sopB, sopE, sopA, siiA, ssaH* and *flhD*

The analysis used in this work consisted of two parts: 1) *de novo* motif detection in a tight list of potential HilA targets (MDscan), 2) filtering of the predicted motifs by selecting those that best explain the change in gene expression (MotifRegressor).

3.3.1 Motif detection: MDscan

The ChIP-chip analysis indicated that there were 209 binding sites for HilA (p-value < 0.05), many of which are probably false positives since the HilA regulon is not expected to contain such a large number of direct targets. Based on the microarray analysis, a group of genes transcriptionally affected by HilA was detected that partially overlapped with the set of genes found to be bound by HilA in the ChIP-chip analysis. The intersection of both gene lists is expected to correspond to the direct and functionally active targets of HilA. A list of 19 genes constituted this intersection (Table 3.1).

Table 3.1 A list of genes that were deferentially expressed in the microarray analysis (|d-value| > 2) and enriched for binding sites in the ChIP-chip analysis (p-value < 0.05)

Gene no.	Gene name	Differential gene expression (d-value)	ChIP-chip enrichment (p-value)
STM1091	sopB	16.3	3.60E-02
SL2674d	sopE	10.8	6.10E-03
STM2066	sopA	10.4	2.60E-02
STM4257	siiA	9.7	2.6E-02f
STM2899	invF	9.6	1.30E-02
STM2874	prgH	7.3	1.60E-02
STM1407	ssaH	–5.9	2.70E-03
STM3203	ygiM	–5.6	1.90E-02
STM3598		5.5	2.70E-02
STM0772	gpmA	3.9	3.60E-02
STM2287	sseL	–3.0	3.70E-02

STM2283	glpT	–2.9	1.60E-02
STM1925	flhD	–2.8	6.00E-04
STM4298	melA	2.6	3.90E-02
STM0670	miaB	2.4	3.10E-03
STM3869	atpF	–2.3	2.50E-03
STM1377	lpp	–2.1	7.40E-03
STM3921	wecC	2	7.60E-04
STM1533		2	6.90E-03

Promoter regions of this sub-selection were subjected to a motif search by MDscan (Liu, Brutlag & Liu, 2002). As mentioned in chapter 2, MDscan is an algorithm for de novo motif detection in a list of ChIP-chip ranked data set. It uses a subset of top ranking genes to look for overrepresented sequences. Groups of similar overrepresented sequences are used to build PWMs, which are in turn used as seeds to scan the remaining sequences and update the motif models. A user-defined number of top scoring motifs are reported.

The algorithm was applied to the 19 genes mentioned above, ranked according to their ChIP-chip scores, searching for motifs between 6 - 20 bp long. The top ten sequences were used to look for seed motifs, and the entire list was used to update the motifs. Whenever found, the top scoring five motifs for each length were reported, making a total of 39 motifs (Table 3.2). The MDscan scores listed in the table are based on MAP scoring function (maximum *a posteriori*), which reflects the degree of over-representation of a motif compared to its expected random occurrence in the sequence. The average probability reported for each motif is the average probability of the most probable bases in each position, and it reflects the degree of conservation of the motif in the aligned sequences. Although the scores can work as guidelines towards defining the correct motif, they are not conclusive. On the one hand the MDscan scores show a clear bias towards shorter motifs, and on the other, the average probability does not reflect the overrepresentation of the motif compared to the background. To define which of these motifs is more likely to be the correct one, the expression data is used more fully as explained in the next section.

Table 3.2 A list of motifs found by MDscan. Whenever found, the top 5 motifs of lengths 6 – 20 are reported. 'Score' refers to MDscan match score. 'Average probability' is the average probability of the most probable base at each position of the motif. The 'Number of aligned sequences' is that of all sites used to build the motif.

Motif number	MDscan score	Average probability	Number of aligned sequences	Consensus
Motif.N2.6.1	3.865	0.793	48	TCCTGA
Motif.N2.6.2	3.703	0.778	58	CAGCAG
Motif.N2.6.3	3.541	0.784	49	TGACGT
Motif.N2.6.4	3.487	0.779	62	GCTGAT
Motif.N2.6.5	3.487	0.754	53	CATCAC
Motif.N2.7.1	3.232	0.745	41	TCCTGAT
Motif.N2.7.2	3.043	0.81	53	TTTTTAC

Motif.N2.7.3	2.998	0.797	38	TGAATTA
Motif.N2.7.4	2.931	0.753	34	TACAGAA
Motif.N2.7.5	2.875	0.789	33	TTACACA
Motif.N2.8.1	2.935	0.699	44	CCTCCTGA
Motif.N2.8.2	2.767	0.697	35	CTGCTGCT
Motif.N2.8.3	2.702	0.721	44	TGCTGTAA
Motif.N2.8.4	2.701	0.709	39	ACGTCACA
Motif.N2.8.5	2.669	0.655	30	CAGTGGGA
Motif.N2.9.1	2.502	0.668	25	GCCTTTCTG
Motif.N2.9.2	2.499	0.657	34	AGCAGCGGG
Motif.N2.9.3	2.429	0.698	34	TACGTCACA
Motif.N2.9.4	2.428	0.662	31	CTGCTGCAG
Motif.N2.9.5	2.348	0.736	48	GCTTTTTTT
Motif.N2.10.1	2.33	0.655	37	CTGCTGCAGA
Motif.N2.10.2	2.246	0.692	39	TTATGTACAT
Motif.N2.10.3	2.22	0.675	31	TGTGCTGTAA
Motif.N2.10.4	2.18	0.666	25	GTTAACAGTA
Motif.N2.10.5	1.999	0.667	22	ACAATATGAG
Motif.N2.11.1	2.092	0.663	25	TTTCATCAGGA
Motif.N2.11.2	2.083	0.691	37	TTATGTACATA
Motif.N2.11.3	2.08	0.657	25	CAGAAGATTAA
Motif.N2.11.4	2.059	0.689	38	AAAATAAGTGC
Motif.N2.11.5	1.967	0.662	25	GATGCTTCTTT
Motif.N2.12.1	1.863	0.653	20	GCGTTTCAGGAA
Motif.N2.13.1	1.809	0.689	30	AAATAAAAACATT
Motif.N2.13.2	1.739	0.662	24	TTTACTTTGTTTA
Motif.N2.13.3	1.68	0.685	25	GATTAATTAATAA
Motif.N2.13.4	1.664	0.67	12	CACATTAAAGTGA
Motif.N2.14.1	1.623	0.658	14	CACATTAAAATGAC
Motif.N2.15.1	1.631	0.666	24	AAAATAAAAACATTA
Motif.N2.16.1	1.66	0.655	12	TGTGATGTACATCACA
Motif.N2.19.1	1.483	0.656	10	AATGTGATGTACATCACAT

3.3.2 MotifRegressor

MotifRegressor has been originally developed to detect motifs in a set of genes that undergo changes in expression in a certain condition (Conlon, Liu, Lieb & Liu, 2003). The algorithm is a combination of the MDscan module and a regression module to fit the expression data (Figure 3.1). Candidate motifs found by MDscan are used to screen a list of genes with an altered expression, and each gene is given a motif score called the sequence motif-matching score, S_{mg}. This score determines how well the upstream sequence of a gene g matches a motif m, in terms of both degree of matching and number of sites. For each detected motif, a simple linear regression is performed between its respective motif scores and the expression values for all the genes in the list according to this model:

$$Y_g = \alpha + \beta_m S_{mg} + \varepsilon_g$$

where Y_g is the expression value of gene g, S_{mg} is the score of motif m in gene g, and ε_g is the gene-specific error term. The baseline expression α and the regression coefficient β_m are estimated from the data. Motifs that can significantly (p-value ≤ 0.01) explain the observed differential expression levels are retained. A step wise multiple linear regression with those selected motifs as variables results in selecting the motif combination (module) that best explains the differential expression experiment (Figure 3.1)

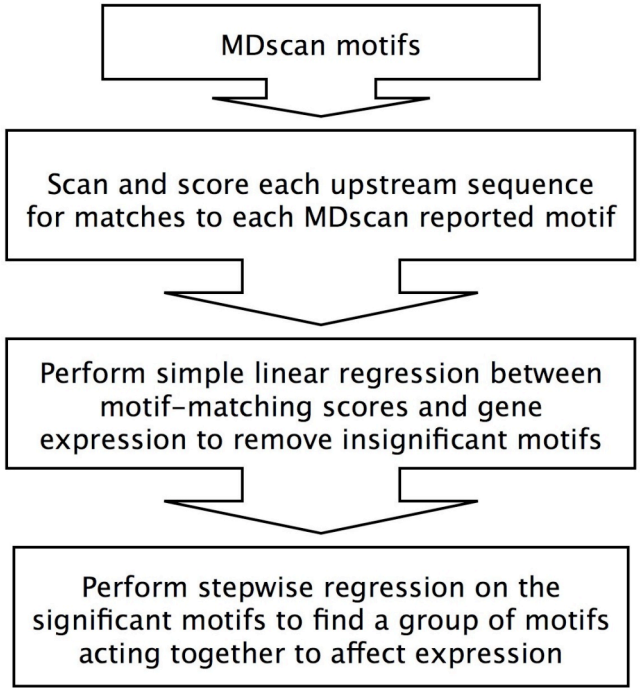


Figure 3.1 MotifRegressor strategy. Adapted from (Conlon, Liu, Lieb & Liu, 2003). We applied the algorithm on the previously detected motifs, and had two motifs that could significantly (p-value < 0.05) explain the expression data (Table 3.3). The last step of the algorithm, i.e. multiple regression, was not performed since our data are strictly HilA analysis data, and only the HilA motif is expected to be present. The higher scoring motif, Motif.N2.11.1, was compared to the previously reported HilA box (De Keersmaecker *et al.*, 2005) and was found to partially overlap with it: CAtCAGga in the reported consensus sequence overlaps with CATCAGGA from Motif.N2.11.1. Thus, this motif was considered for further analysis.

Table 3.3 Candidate motifs that best explain the expression data according to MotifRegressor. P-value: for the correlation coefficient; S-Coeff: the simple linear regression coefficient value; R-sq: Pearson's correlation coefficient of regression

Motif	Consensus	P-val	S-Coeff	R-sq
Motif.N2.11.1	TTTCATCAGGA	2.29E-03	6.71E-01	4.50E-01
Motif.N2.10.1	CTGCTGCAGA	8.72E-03	5.98E-01	3.58E-01

The motif logo is shown in Figure 3.2. Several occurrences of this motif were reported for each gene as shown in Table 3.4. For the two genes in which the HilA motif was previously delineated, namely *invF* and *prgH*, some of the sites match those reported in De Keersmaecker *et al*. (highlighted in Table 3.4).



Figure 3.2 HilA motif logo based on the results of MDscan. The motif was generated by http://weblogo.berkeley.edu/logo.cgi

Table 3.4 Candidate HilA motif sites in target genes. Gene numbers and names are shown for each gene in which a motif site(s) was predicted. The strand, locations in the promoter sequences used, and sites of the predicted motif are shown. Previously reported sites are highlighted (De Keersmaecker *et al.*, 2005).

Gene number	Gene name	Strand	Location	Site
STM2899	invF	+	220	TTAAATCAGCA
STM2899	invF	-	166	CTGCAGCAGCA
STM2899	invF	+	285	TTGCATCAGGA
STM2899	invF	-	142	TGTCATCTGGA
STM2899	invF	-	116	TTATATCAGGA
STM4298	melA	+	216	TTTCACCGGGA
STM2874	prgH	+	243	TTTCATCAGGA
STM2874	prgH	-	44	TTTCAGTAGGA
STM4257	siiA	+	254	TCTCATCTGCA
STM4257	siiA	-	712	AGTCATCAGCA
STM4257	siiA	+	583	TATCCTCTGGA
STM4257	siiA	-	415	TATCTTCAGCA
STM4257	siiA	+	930	TGGTAGCAGGA
STM4257	siiA	-	63	TTACACCAGGC
STM4257	siiA	-	51	TTTCAGCAGAC
STM4257	siiA	+	1025	TAATATCAGGA
STM1091	sopB	+	39	TTTACGCAGGA
STM1091	sopB	-	245	ATGCACCAGGA
STM1091	sopB	-	215	TTTAATGAGGA
SL2674d	sopE	+	94	TTGAATCAGCC
SL2674d	sopE	-	27	TTCGCTCAGGA
STM1407	ssaH	-	28	TTCCAGCAGCA
STM1407	ssaH	+	27	TTGCTGCTGGA
STM2287	sseL	+	56	TTCCATCAGAC
STM2287	sseL	-	329	TTAAATCAGGC

However, the list of genes used for the analysis did not contain two previously known targets of HilA: *sicA* and *hilA,* which were confirmed to bind by the macroarray analysis. We performed a motif scan on the promoter regions of the genes with Motif.N2.11.1. The scan revealed five motif sites in *hilA* (Table 3.5), but non in *sicA*. In this case too, the previously reported HilA binding site was matched here (highlighted in Table 3.5). An alignment of motif sites that matched previously reported ones, and all motif sites for the rest of the genes in the positive strand is shown in Figure 3.3.

Table 3.5 Motif sites found in the *hilA* promoter

Gene number	Gene name	Strand	Location	Site
STM2876	hilA	+	31	CTCTATCAGGA
STM2876	hilA	-	611	TTACACCATCA
STM2876	hilA	-	367	CGATATCAGGA
STM2876	hilA	+	764	CTGCACCAGGA
STM2876	hilA	-	203	TGTCATCTGGA

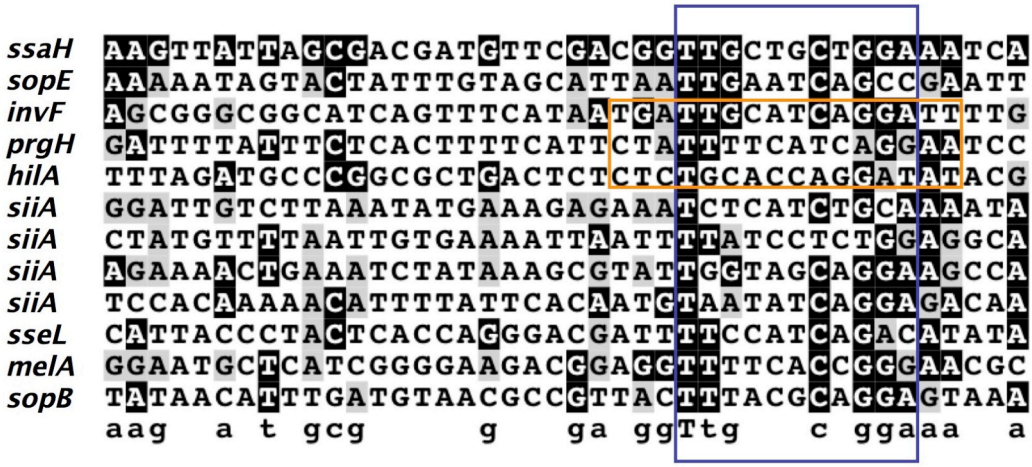


Figure 3.3 Alignment of the putative HilA motif sites. Our analysis revealed several sites for the HilA motif (tables 1 and 2). Here we show all sites found on the positive strand, except for genes with a previously known binding site, we only show the sites that coincided with the known sites. The HilA-box is indicated with a blue box; previously described sites are enclosed in an orange box; consensus sequence is shown at the bottom

3.3.3 Discussion

In this work we have identified 9 genes as direct targets of the HilA type III secretions system in the pathogenic bacterium *S. enterica* serovar Typhimurium by finding a HilA motif in their promoter regions. We have also suggested some potential binding sites that can be further investigated by wet lab experiments. Some of the targets revealed by our

study have been already known (*prgH, invF*, *hilA*) and the exact binding sites reported for them matched those predicted here. Other targets (*sopB, sopE, siiA, ssah, sseL, melA, flhD*) are a suggested extension to the HilA regulon and have been shown to be involved in invasion and enteritis (*sopB* and *sopE,*), secretion (*ssaH*), motility (*flhD*), adhesion to epithelial cells (*siiA* operon), and other processes related to pathogenicity (Thijs *et al*. 2007).

The way we applied MotifRegressor constitutes a novel and more conceptually proper approach, in our opinion, towards the discovery of motifs. The original implementation of MotifRegressor used expression data as the sole source of data: MDscan discovered potential motifs from a list of genes ranked according to their change in expression, and the expression was used once more for regression. Motifs found originally in the 'seed' list are more likely to better fit the expression, and thus there is circularity in the method. In our application, the candidate motifs of MDscan come out from a set of ChIP-chip top scoring genes, which is a more biologically relevant approach since the ChIP-chip assay is an assay of a TF's binding to its motif sites. Then the candidate motifs were fitted to a different set of data i.e. expression, hence avoiding redundancy in the data sources and adding confidence to the results.

Although the method of regressing expression values on motif-scores was not successful for an extended list of ChIP-chip enriched genes (data not shown), the method proved useful for a reduced list of highly potential targets. Probably the signal to noise ratio was too high to detect the true motifs in an extended list. Indeed, many of the genes enriched in the ChIP-chip experiment are not expected to be true positives considering the known involvement of HilA in virulence, a rather non-global cellular process. The binding of HilA to 'false positives' could be due to the likely presence of non-functional binding sites in their promoters.

Chapter 4

4 Detection of de novo motifs in *Bacillus subtilis* by Phylogenetic footprinting

4.1 Introduction

In this chapter we describe the prediction of novel motifs at a genome level in the Gram-positive model organism *B. subtilis*, so use them in the inference of the transcriptional network in the next chapter. Studying the transcriptional network of *B. subtilis* is interesting as the organism is capable of surviving in many diverse environments, and undergoes a complex condition-dependent differentiation process, i.e. sporulation.

The method employed to discover motifs in this organism, is based on phylogenetic footprinting and the concept of co-regulation. Phylogenetic footprinting was introduced in chapter 2 as one of the methods for *de novo* motif detection. As a reminder, the principle underlying the method is that functional elements in non-coding sequences are conserved over evolutionary times relative to the non-functional landscape. Thus, upon comparison of different species, conserved regions are seen as functional footprints. A proof of concept was demonstrated by the enrichment of regulatory elements in conserved regions between human and mouse intergenic sequences (Levy, Hannenhalli & Workman, 2001; Wasserman, Palumbo, Thompson, Fickett & Lawrence, 2000). Wasserman *et al*. have found that 98% of the known muscle regulatory motifs are situated in 19% of the most conserved human-mouse regions. Phylogenetic footprinting has been initially performed on certain genes of interest and was successful in identifying regulatory elements for many genes, for example ε-globin and γ-globin (Tagle *et al*., 1988), rbcL (Manen, Savolainen & Simon, 1994), cystic fibrosis transmembrane conductance regulator (Vuillaumier *et al*., 1997), tumor necrosis factor-α (Leung et al., 2000), and interleukin(IL)-4, IL-13, and IL-5 (Loots *et al*., 2000). The method was applied on a genome-wide level for the first time by McCue *et al*. analyzing intergenic regions of six proteobacterial species (McCue *et al*., 2001). Genome-wide application as such results in a large set of potential motifs (2097 sites were identified in the McCue study). For this reason, we developed a strategy to reduce the output of phylogenetic footprinting based on co-regulation.

In this work we employ an approach similar to the one we described earlier (Monsieurs et al., 2006), in order to detect motifs in the *B. subtilis* genome. We validated our results with expression data and with functional enrichments.

4.2 Materials and Methods

4.2.1 Data set

Four Bacillus species were selected for the phylogenetic footprinting step, based on their evolutionary spacing. The following genome sequences were downloaded from GenBank: *B. subtilis* (AL009126), *B. anthracis* (NC_007530), *B. licheniformis* (NC_006322) and *B. cereus* (NC_004722). Orthologous gene sets were identified by pair-wise reciprocal BLAST best hits (Zhang, Schwartz, Wagner & Miller, 2000). The entire non-coding sequence, upstream of the start codon on the genomic plus strand i.e. the strand with the same orientation as the coding sequence of the gene, was extracted. To minimize the noise in the data set during motif detection, we only retained gene families that contained an ortholog in all of the four species and had an upstream sequence of at least 40 bp. As such a total of 1284 gene sets were selected, resulting in 1284 sets of intergenic sequences for the corresponding genes.

4.2.2 Retrieving conserved blocks (step 1 in Figure 4.1)

For the phylogenetic footprinting step, we used BlockSampler, a Gibbs sampling algorithm optimized for prokaryotic sequences (Monsieurs *et al.*, 2006). BlockSampler delineates conserved regions (or blocks) in each of the orthologous datasets. *B. subtilis* was used as the reference sequence for BlockSampler. We used species-specific third-order background models based on the intergenic sequences of each of the four species. The algorithm was run 100 times per orthologous dataset with the default parameters, resulting in a list of 100 conserved regions (blocks) per gene set. Overlapping blocks for more than 75% were filtered out, by keeping only representative blocks for each overlapping group, which had the best log likelihood scores. For each dataset, a maximum of 6 best scoring, non-overlapping blocks were retained.

4.2.3 Pair-wise comparison of blocks (step 2 in Figure 4.1)

Each conserved block was converted into a PWM, and all of the resulting PWMs were compared pair-wise using the Kullback-Liebler distance measure implemented in the algorithm BlockAligner (Monsieurs *et al.*, 2006). Blocks exhibiting a minimal overlap of 6 bp were reported. Scores (p-values) were assigned to each alignment by performing 100

randomizations of the dataset. Alignments with a p-value < 0.0001 were considered significant and retained for cluster analysis.

4.2.4 Clustering of Blocks (step 3 in Figure 4.1)

The scores of the pair-wise alignments generated in the previous step (i.e. p-values) were used as input for a graph-based fuzzy clustering algorithm (Joshi, Van de Peer & Michoel, 2008), in order to delineate clusters or groups of similar blocks derived from different gene sets. Ten clustering solutions were obtained and merged, and a probability cut off value of 0.5 was used to obtain tight clusters. Cluster analysis resulted in 300 clusters containing 2–12 blocks per cluster.

4.2.5 Extraction of motifs (step 4 in Figure 4.1)

In this final step a consensus motif was extracted from each of the obtained clusters. To this end, the reference sequence from *B. subtilis* was extracted for each block within a cluster. This results in sets of unaligned *B. subtilis* intergenic sequences, corresponding to the clusters. Each set contains sequences derived from at least two different *B. subtilis* genes and containing a conserved motif. This conserved motif was extracted by locally aligning these reference sequences using (MotifSampler; Thijs *et al.*, 2002; INCLUSive-3.1.2 modified by Claeys *et al.*). The algorithm was run with the following parameters: prior = 0.5, runs = 100, length = 6, 8, 10, 12 and 14. Post-processing was performed as implemented in MotifRanking (INCLUSive-3.1.2): for each dataset, a motif was selected that had the highest log likelihood score and that occurred most frequently in the result set (most obvious local optimum). 160 motifs were found ultimately.

4.2.6 Screening the *B. subtilis* genome

The method of Hertzberg (Hertzberg, Zuk, Getz & Domany, 2005) (Hertzberg et al., 2005) was used to screen the genome of *B. subtilis* with the found motif models. For this step, the intergenic sequences of *B. subtilis* (as described in section 4.2.1) were extended by an additional 100 bp downstream of the translation start site.

4.2.7 Calculating functional enrichment and overlap with targets of known regulators

The functional categories for all genes downloaded from DBTBS. Enrichment of a group of genes for a specific functional category was calculated by means of the hypergeometric distribution. Target genes of known regulators were obtained from DBTBS as

well, and the enrichment of a certain cluster for targets of a specific regulator was calculated by the hypergeometric distribution.

4.2.8 Associating *de novo* motifs with potential regulators

A full list of regulators in *B. subtilis* that do not have a documented regulatory motif was compiled by merging the known regulators from DBTBS for which no motif was identified, with the hypothetical regulators in *B. subtilis* retrieved from the DNA-binding domain (DBD) database (Kummerfeld & Teichmann, 2006). Regulators were assigned to their potential motifs based on the principle of the "conserved neighborhood of a regulator and its target genes in bacteria": for each of the modules regulated by one of the novel motifs, we assigned one or more regulators from the above mentioned list of regulators, that were 'sufficiently close' to at least one of the target genes in the module. A measure of 'sufficiently close' was defined as follows: for each known regulator in *B. subtilis,* we identified its closest documented target gene (distance measured in base pairs). The median of those 'closest distances' for all known regulators was used as an estimate for the minimal distance within which a regulator and its corresponding closest target gene are co-localized.

4.3 Results

4.3.1 General strategy

As most motif detection methods suffer from high rates of false positives (Wasserman & Sandelin, 2004), reduction of the search space is necessary to increase their efficiency. Here, we rely on phylogenetic footprinting and the concept of co-regulation as means for data reduction. Our strategy (summarized in Figure 4.1) comprises several steps. Initially, several species are used to do the footprint (step 1 in Figure 4.1). Stretches of conserved regions are called blocks. Since most TFs regulate more than one gene, those blocks are expected to cluster around true motifs. For this purpose, the blocks are compared with an alignment algorithm (step 2 in Figure 4.1), and subsequently clustered (step 3 in Figure 4.1). From each of the found clustered, a motif is derived (step 4 in Figure 4.1). The process is described in detail in the following section.

Four Bacillus species were selected for phylogenetic footprinting: *B. subtilis*, *B. anthracis*, *B. licheniformis* and *B. cereus*. The species show moderate evolutionary distances reflected in the sequence identities from the average pair-wise alignments of their intergenic regions (Table 4.1). At these evolutionary distances, the nonfunctional landscape have

diverged enough to allow many functional sequence signals to stand out from the "noise," but the sequences retain enough overall similarity to enable their alignment (Cliften *et al*., 2003). This was echoed by the moderately sized conserved regions that were found in the subsequent footprinting step. Four-way orthologous genes were identified (1284 genes) and their upstream intergenic regions were subjected to footprinting.

For the purpose of revealing the conserved regions (blocks) in these sets of sequences, we used a Gibbs sampling based algorithm called BlockSampler (Monsieurs *et al*., 2006). The algorithm revealed several blocks for many sets of sequences. Thus, a filtering step was done keeping only the best scoring blocks, with a maximum of six blocks per set (see materials and methods). Each block was translated into a PSSM. PSSM's of all of the blocks were compared pair-wise, using the alignment algorithm BlockAligner (Monsieurs *et al*., 2006). The algorithm allows for a statistical assessment of the comparison results in the form of p-values. These values were used to cluster the blocks with a fuzzy clustering algorithm (Joshi et al., 2008). Each cluster contains sequences in which a motif is more likely to exist. We used the *de novo* motif detection tool, MotifSampler (Thijs *et al*., 2002), to search for motifs in the *B. subtilis* sequences of each cluster, and found 160 motifs in total, ranging in size between 6 to 14 bp.

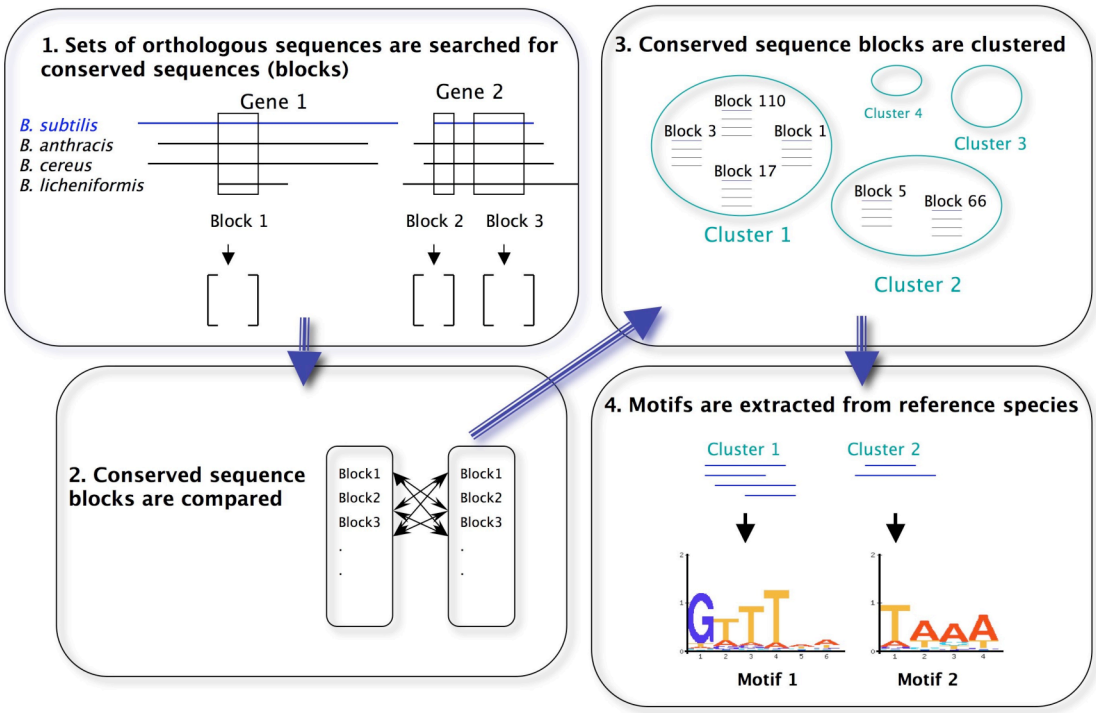


Figure 4.1 Workflow of the de *novo* motif detection. 1) Search for motifs in the orthologous space: conserved sequences (blocks) are extracted from the orthologous intergenic regions of four Bacillus species by means of BlockSampler, a Gibbs-sampling algorithm. The sequence of *B. subtilis* is indicated in blue. Detected conserved blocks are represented as PWMs; 2) Search for motifs in the coexpression space: All PWMs are aligned pair-wise and similarity scores are obtained 3) Similarity scores are used as input for clustering in order to identify groups of sequences sharing a common motif (or block as represented by a PWM); 4) Each block in a cluster corresponds to a different motif site of one particular regulator, conserved across the four different species. From each block within a cluster the sequence of *B. subtilis* is used to construct a *B. subtilis* specific motif model

Table 4.1 Species used for the footprint analysis. The sizes of their genomes are given in kilo bases (Genome Size) and the number of protein coding genes (No of Genes). The pair-wise phylogenetic distance between the reference species, *B. subtilis,* and each of the other three species is reflected in terms of the shared number of orthologous genes, expressed as an absolute number and as a percentage of the number of genes of *B. subtilis* (No of orthologs shared with *B. subtilis*), and the overall pair-wise sequence identity scores (Itergenic sequenceID with *B. subtilis*), calculated as the average score of all Smith-Waterman multiple alignments between the upstream sequences of the orthologous gene pairs.

Species	Genome size (kb)	No of genes	No of orthologs shared with *B. subtilis*	Itergenic sequenceID with *B. subtilis*
B. subtilis	4215	4105	-	-
B. anthracis	5227	5309	2419 (59%)	52.3%
B. licheniformis	4223	4196	3055 (74%)	63.0%
B. cereus	5412	5234	2415 (59%)	52.0%

4.3.2 Benchmarking against DBTBS

In order to validate our results, we compared them to the previously known motifs, as reported in DBTBS. In

Table 4.2 the entire list of the motifs retrieved by our analysis is displayed. For each putative motif the consensus sequence is shown together with genes of *B. subtilis* that contributed to the detection of the respective motif. For each motif, we assessed whether the genes contributing to them are enriched for certain functions (see materials and methods). Enrichment for a certain biological function would indicate that the cluster of genes is meaningful, thus, more likely to similarly regulated. This proved to be true for some of the motifs

Table 4.2).

To calculate the overlap with known motifs in DBTBS, we assessed whether the gene sets contributing to the motif were enriched for targets of known regulators listed in DBTBS. If an overlapping regulator was found, we manually verified the predicted consensus sequence with the documented one. However, this method is highly unreliable because it the motif representation depends to a high extent on number and content of the constituting sequences; most of the motifs reported in DBTBS are based on many sequences. We were able to verify few motifs, e.g motif_163 is similar to the motif of ResD. The genes of motif_163 are enriched for electron transport chain and ATP synthesis, functions that are compatible to the known role of ResD in respiration.

Table 4.2 List of predicted motifs. For each motif predicted by in this study the consensus sequences is shown, the genes contributing to the motif, the functions for which these genes were enriched, and the known regulators that have a significant overlap of their target genes with the genes of our motifs.

Predicted motif	Consensus	Genes	Functional enrichment	Overlapping TFs
motif_10	CnnGnGnATATGGC	pbuG purE xpt	• Transport/binding proteins and lipoproteins • Metabolism of nucleotides and nucleic acids	PurR
motif_102	TTATCC	recO yqeN		
motif_103	TAAwTGAyAAGyAm	lysC ydzA		
motif_105	GsTTTTTT	gltX uvrB aroK gerM	• Germination • DNA restriction/modification and repair • (Protein synthesis) Aminoacyl-tRNA synthetases	LexA SpoIIID
motif_106	kkATTATrATAAAA	dacA yqxD		SigB
motif_107	kAwyAyACACTATT	ydiF speA		
motif_108	wnTnnTwGTnnA	pdhA ykuJ yjbI	• (Metabolism of carbohydrates and related molecules) Main glycolytic pathways	
motif_11	TGATrGwCyATCAT	ykrU ykrV	•	
motif_112	GwwTGAGG	ytpA yozE	•	
motif_114	TTnwnwrGGA	sleB mlpA yisN	• Germination • Protein modification	SigG
motif_115	AGnnnTTTnTnw	ywbO yviA yhbB		
motif_116	TAyAATArCT	tagO yusZ		
motif_117	TAnAATnn	yjbX ykvW glmS		PerR
motif_119	CTGCGT	leuS alaS	• (Protein synthesis) Aminoacyl-tRNA synthetases	
motif_12	TGAnAAnnATTnTC	fer fhuD ywjA ypbB ykuN dhbA		Fur
motif_122	CsyTGsGCATAwwA	ykoY yitB	• Metabolism of sulfur	SigK
motif_123	AGGmAG	ribA iolS	• Metabolism of coenzymes and prosthetic groups	
motif_124	wrAswAsAykAC	trxA yqhV	• Membrane bioenergetics (electron transport chain and ATP synthase)	CtsR SigB YlbO
motif_125	GmAGCGAAGGwsCy	yusU yphE		
motif_128	nAAnwAnGGn	mbl ytkL dacA	• Cell wall	SigB
motif_13	ATCrAATTyGATmT	yutJ yuzB		
motif_133	TTGmAmCGkGAyww	ywpE ywzC		
motif_134	ATCGTA	gyrA tuaD	• Cell wall • DNA packaging and segregation	
motif_135	TkATAACGAT	yrkD yfmM		
motif_137	AAATACAAAwwAAC	yqiX citM		CcpA CitT
motif_138	yAAGrTGG	ycsF rplX	• Detoxification	KipR TnrA
motif_14	yAAAAmAGGAGrmr	yneS yozD		
motif_140	AAGmGGTG	yugE yrhI		AbrB YrhI
motif_141	GTAnnAmGnn	spoIIIAA yvlA ysgA	• RNA modification	AbrB SigW SpoIIID
motif_142	GGArGC	yubB ung	• DNA restriction/modification	

			and repair	
motif_144	TATGAT	yvyH yuiF		
motif_147	GGrAGC	ydbR ssuA		
motif_15	wGTArATyTACw	yktA speA		
motif_154	ACwATCwGATTkyk	ycgM hxlR		
motif_16	sykTTGAAGTCT	yefA yutI		
motif_160	rAGAGm	yvaB yhcW	• Membrane bioenergetics (electron transport chain and ATP synthase)	SigG
motif_163	TAAyTTTkTsAC	ugtP hmp	• Membrane bioenergetics (electron transport chain and ATP synthase)	ArfM ResD
motif_172	mGCTCGAA	ytsP rocD		AhrC RocR SigL
motif_177	TwCCTTTyTT	yabM licB	• Transport/binding proteins and lipoproteins	CcpA LicR
motif_178	AsACAACAyT	comEB yqzE	• Transformation/competence	
motif_179	GmrAyCTyCATG	ykfA yfkK	•	
motif_181	AAATArwGTATA	valS yqzG	• (Protein synthesis) Aminoacyl-tRNA synthetases	SigF
motif_182	CwTrCTGA	yccC ytvI		SigE TnrA
motif_185	AATGTTGr	yefA yuxO		
motif_187	AGwmrATTTATTTk	ykbA ndhF		
motif_19	AAATCryryGATTT	yutD yutC		
motif_192	AwksAGCT	yhaL gerD	• Germination	SigG SpoVT
motif_195	CAwrCTAT	yrhL cheR		
motif_196	AGCAAT	bkdAA yozD	• Metabolism of lipids	
motif_197	TTCkAkGArwTTCk	ycnE yyxA	• Adaptation to atypical conditions	
motif_2	wnAGGAGGTG	rplO purA rpsF rplC yhdP rplJ rpsM divIVA yhbH	• (Protein synthesis) Ribosomal proteins	PhoP PurR
motif_201	GGAGAG	ywmB veg	• Miscellaneous	
motif_210	GCTGrA	yocS yjbD		PerR SigM
motif_214	ATrATAAC	yqgY yttB	• Transport/binding proteins and lipoproteins	
motif_215	nGAnrnGAAA	yteA yqjN recU	• DNA recombination	SigG
motif_218	krTCAAwTGATTCy	thiD penP	• Metabolism of coenzymes and prosthetic groups	
motif_219	GGGCAGrm	yojG yjbG		
motif_220	GATATGCm	yvfV ytaF		
motif_225	CTTwATTGwCCrrA	ylpC clpP	• Adaptation to atypical conditions	CtsR SigB
motif_228	CrTGCA	comEA comER	• Transformation/competence	ComK
motif_23	AkATkGkmCmATmT	yvdD yvdC		
motif_230	TyCkCkGAATGArG	yyaC yusW		SigF
motif_231	GCAGGT	secDF infC	• (Protein synthesis) Initiation	
motif_232	TyCTTTyTTTTAwG	yutI hepS		
motif_233	CyCyCAmC	ybbK yogA		
motif_234	GGCCAA	ydiF yefA		
motif_235	TTGAskTsAA	mrpA ackA	• Transport/binding proteins and lipoproteins	CcpA
motif_240	sTmCrTCTTG	rplK ywaD		
motif_243	TTrTATkATArTAA	yqfN ykuJ		
motif_244	rrCTTkAAGk	yaaO ccdA	• Membrane bioenergetics	

			(electron transport chain and ATP synthase)	
motif_245	ATyATATTGA	dnaE yoeD	• DNA replication	LexA
motif_246	AATGwAryGwATTG	yjbJ ftsE	• Phage-related functions	SigD
motif_247	AAAGnnGw	rnhC ykuR yunE	• DNA replication	
motif_248	GGrAwATG	etfB ysxE	• Membrane bioenergetics (electron transport chain and ATP synthase)	
motif_250	CsTCTCrTsTTT	uppS ylbG		
motif_252	nnGATG	rplS yjhB spoVAA		SigG SpoVT
motif_254	TTkTCA	hom yqzG		SigF
motif_256	yTGTGAmGysTCAT	ytkA sodA		
motif_257	CyCmTTTATAAAkG	yfkA yfjT		
motif_258	GwTCrwAAGGAGrA	yplQ sdaAB	• Miscellaneous	
motif_259	TTAkTGATAA	ytlP dat		
motif_260	TnnTnTTTTnnn	yugP yitS yumB		
motif_261	kGGAATAT	arsC yphA	• Detoxification	SigF
motif_263	TAAAGn	yojE adhA ytcG		
motif_264	AkTCswwsCGGA	sigA yusL	• Metabolism of lipids • (RNA synthesis) Initiation	SigD YvbA
motif_265	CTTGAT	yqjM nusB	• Membrane bioenergetics (electron transport chain and ATP synthase) • (RNA synthesis) Termination	PurR
motif_266	TAnnrnrArG	yrkI sigI uppS	• (RNA synthesis) Initiation	SigI
motif_267	AmAATATGAA	ytzC ypdA		
motif_268	TnnTnnnmTTnA	yshA yxiE yitJ		
motif_269	ACAwTTmACw	ydaH yitL		
motif_271	GTnnnCnn	rpoA yqhP relA	• Metabolism of nucleotides and nucleic acids	SigB
motif_274	ArGkTGCAGAky	ykfA ypbR		Fur
motif_275	TmGkAAAGkA	ytkK tepA		
motif_276	TGTskm	metE cysK		YrzC
motif_277	GnGnnnnTnwGA	ileS yumC licB		CcpA LicR
motif_278	AAkAACAATGAA	ezrA leuS	• Cell division (Protein synthesis) Aminoacyl-tRNA synthetases	
motif_281	nTnTTnnTAnrA	hmp lspA divIB	• Membrane bioenergetics (electron transport chain and ATP synthase) • Cell division	ArfM ResD
motif_282	CkTCwwCTCTATCA	glpF rho	• Transport/binding proteins and lipoproteins	CcpA GlpP
motif_284	TCmCATGATswkrs	yphA pstS		PhoP SigF
motif_285	TTATTwTGrAyAGC	ylaI ydaH		
motif_288	GmCAArGTGyAAAA	rocR ytkK		RocR
motif_289	ArGwkGTAyATmCr	tlp treP		CcpA TreR
motif_29	TnTTwCnTnT	yycA ysfC yphE		
motif_291	TCATTAwwCAkA	proI yvgZ		
motif_293	TGATrGGAAArs	ytxG yhfE	• Adaptation to atypical conditions	SigB SigH
motif_294	ATkTGGTA	yesE yqzG	• No similarity	SigF
motif_295	AwGTTTAATs	ypuI pfkA	• (Metabolism of carbohydrates and related molecules) • Main glycolytic pathways	
motif_298	GGCGGm	yrhF ffh	• Protein secretion	

motif_3	AGGGTGGwACCrCG	ileS tyrS cysE hisS valS thrZ alaS proB thrS	• (Protein synthesis) Aminoacyl-tRNA synthetases	
motif_303	ArrTTAsyTGGAkG	hxlR murE	• Cell wall	
motif_306	CGCAGG	drm ywnE	• Metabolism of nucleotides and nucleic acids	
motif_310	ATAnwATnnnwnwn	yqhM ybfQ yqhL		
motif_34	TAkAATAArG	padR yitL		
motif_36	nGGAnnnTnTTT	lonA ytxK ycdC	• Adaptation to atypical conditions	CtsR
motif_37	nGTnAC	ydiB yrrT yrzA		YrzC
motif_39	TCACATsG	yutG opuAA		
motif_4	TGnTATAATnnmnn	sinR mutS yrzC	• DNA restriction/modification and repair	
motif_40	GnAGGnTn	ypmA rplJ pheT	• (Protein synthesis) Aminoacyl-tRNA synthetases	
motif_41	ArTCCAGAGAGGnT	pyrR pyrP pyrB	• Transport/binding proteins and lipoproteins • Metabolism of nucleotides and nucleic acids	PurR PyrR
motif_44	AAATTATAATTT	yslB sdhC	• (Metabolism of carbohydrates and related molecules) TCA cycle	
motif_45	TTTAsCwswGsTAA	yloU rpmB		
motif_46	AmyGTACrkT	yrzC yrvN		
motif_49	TATTATAATAmA	yrhF yrhE		
motif_5	CAAwnTGTGGAT	yqeG yydA dnaA ywlC rpmH	• DNA replication • (Protein synthesis) Ribosomal proteins	DnaA
motif_50	TGkGCmCA	yqhR yqhQ		SigB SigF
motif_52	AkmTGATwwwATCA	ansR ansA		AnsR
motif_53	AAAATwAAAATrAA	ppiB ypuA	• Protein modification	
motif_57	rwrTTAGCTAAywy	ypjD ypjC		
motif_58	AAGGGGAkGswG	yutE appB	• Transport/binding proteins and lipoproteins	
motif_6	wnCTTGACAn	ymdA ylxM ypfD guaB		
motif_60	TAATTGTCAG	ndk ymfC ywpB nifZ		
motif_61	TTGGArwT	lipA kbaA		
motif_62	CAArTkmATwATAT	yktA speA		
motif_63	AAyrkTTGCAAm	yloU rpmB		
motif_64	CCwyCATsAyGACA	ypjP yvfR		
motif_65	TnnTAA	rpoB yuzB yutJ	• (RNA synthesis) Elongation	
motif_68	TTTwTAwGTArrCC	katE citZ	• (Metabolism of carbohydrates and related molecules) TCA cycle • Detoxification	CcpA CcpC SigB
motif_7	nGTAGCTCAGnTGG	spoIIE yycF glgB		Spo0A
motif_71	nGAnnTAA	ytzH yyaQ ytmQ		
motif_73	AAAAnCnnnnTnwn	yrhA ylxA ydfL		
motif_75	AAAAAwCwTTTTwk	hmp abrB icd ypfD	• Membrane bioenergetics (electron transport chain and ATP synthase) • (Metabolism of carbohydrates and related molecules) TCA cycle	ArfM ResD Spo0A

motif_76	GCGTyTTyATAGAA	yjbB yutI		
motif_8	GAGGGACnnGCCCn	metK pckA ykrT yjcI ykrW yitJ	• (Metabolism of carbohydrates and related molecules) Main glycolytic pathways • Metabolism of amino acids and related molecules	
motif_81	ATwGCwAGkmACrG	med ytkD	• Transformation/competence	SigF
motif_82	GmTTwCAwACmC	ypcP mtrA	• Metabolism of coenzymes and prosthetic groups • DNA replication	
motif_83	nnnnnTTnnAAmAn	nucA yybE speE	• Metabolism of nucleotides and nucleic acids	ComK
motif_84	nnATGGnAwr	yfhS yrrT qoxA		YrzC
motif_85	wrGrTG	comC ytvB ymcB yeaC	• Transformation/competence	ComK
motif_86	AnCATA	ylxY divIB ruvA	• Cell division • DNA recombination	LexA
motif_88	ACTwkCAATGAATC	yoxA ylbB		SigB SigF SigH
motif_89	nTAGAATG	lytS yubA ytqA spoIIB		
motif_9	GwGyGCwm	ydeP yfkO groES ydiL hrcA	• Protein folding	HrcA
motif_91	krTCGAGTGG	spoIIID med	• Transformation/competence	
motif_92	CwGAAA	drm yurZ	• Metabolism of nucleotides and nucleic acids	
motif_93	TAAwyAGAsAAAGA	ytoA ykuI		
motif_95	GAwTryACGT	pdhA tagO	• (Metabolism of carbohydrates and related molecules) Main glycolytic pathways	
motif_96	AAAyAGATACATAT	yvyD rpoE	• (RNA synthesis) Initiation • (RNA synthesis) Elongation	SigB SigH
motif_97	GnAATG	ycgO yqfZ sspH		SigG SpoIIID
motif_98	kAwAAGyTTwTTGT	dhaS yumC		
motif_99	nCnrAnnrTT	yvgT ywdL rnmV	• RNA modification	

4.3.3 Validation by data integration

There is no straightforward approach to validate the predicted motifs. Typically, when a new motif is predicted in a gene, its binding factor is known and any of the available biological assays e.g. electrophoretic mobility shift assay (EMSA), can be used to confirm the presence and functionality of the motif. Predicting motifs on a genome-wide scale with phylogenetic footprinting comprises difficulty in validation because of the high number of predictions (it is hard to test them all in the wet lab) and the complete lack of knowledge about their corresponding TFs. However by exploiting publicly available data sources such as microarrays, we can filter the most promising predictions from our full list. As TF-gene regulation is conditioned in most cases by the different environmental circumstances of the cell, co-regulation amongst target genes is expected in only a subset of microarrays. For this

reason, we used a bi-clustering algorithm called DISTILLER (Lemmens *et al*., 2009) to validate our motifs. The full application of DISTILLER is discussed in chapter 5.

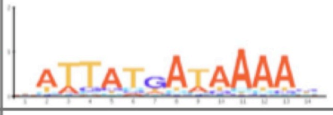
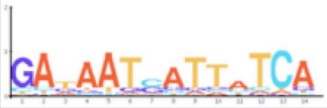
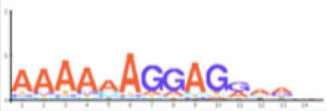
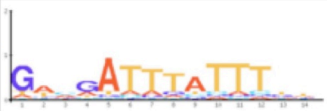
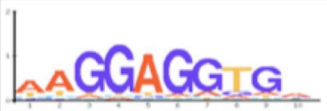
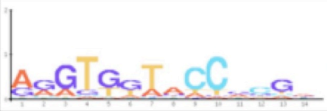
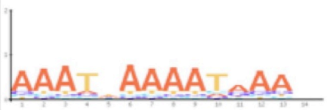
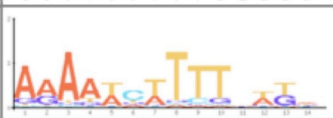
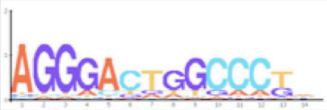
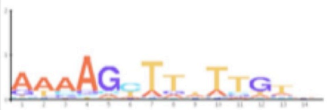
Since the use of phylogenetic footprinting limits the search to only orthologous genes, we wanted to recover occurrences of the motifs in the rest of the genes of *B. subtilis*. Therefore, the predicted motifs were used to screen the entire genome of *B. subtilis* using the method of Hertzberg (see materials and methods). The screening results were used as input for DISTILLER. The data integration tool reported 142 transcriptional modules, each comprising a set of genes co-expressed under a subset of conditions, and containing common motif(s). 13 of our predicted motifs were found to regulate at least one of these modules, indicating their functionality. A list of those motifs with their logos and functional enrichment can be found in Table 4.3. Most of the genes in these modules have no function assigned yet, indicating that they are still largely understudied and could potentially be regulated by these uncharacterized regulators.

To associate these *de novo* motifs with their putative regulators, we relied on two known properties of bacterial networks. First, it was observed that a regulator's gene has a distance constraint with its closest binding site on the genome (Tan, McCue & Stormo, 2005), this means that at least one of the target genes of the regulator is located in the close genomic neighborhood of the regulator's encoding gene. Accordingly, we assigned one or more potential regulators to our predicted motifs based on the presence of the regulator's gene within a minimum distance from at least one of the genes containing the motif. The minimum distance was determined based on the observed regulator-'closest-target' distances of known regulator-target pairs in the *B. subtilis* genome (see materials and methods). Secondly, bacterial regulators are often auto-regulated: 55% of the regulators in *E. coli* and 42% in *B. subtilis* are known to be auto-regulated (Martínez-Antonio & Collado-Vides, 2003) and this study, respectively). If a regulator belongs to a module of a *de novo* motif, it is considered as a candidate for recognizing the corresponding motif and regulating the module. Auto-regulators would also be the closest targets to their own genes and thus fulfill the "close genomic neighborhood" condition mentioned earlier.

Most of the assigned potential regulators (Table 4.3) do not have a known function. However, in some of the cases where the regulator's function is known, it was found to be relevant to the function for which the corresponding module(s) is enriched. For example, one of the assigned regulators of motif-53, AzlB, is a regulator of genes encoding branched-chain amino acid transporters (Belitsky, Gustafsson, Sonenshein & Von Wachenfeldt, 1997). Amino acid transport is the process by which bacteria import amino acids from the environment to be used in building proteins, or to be catabolized (Burkovski, 2002). Thus, there is a need for coordination between the processes of amino acid transport and

metabolism. The latter process is the function for which the modules of motif-53 are enriched, and thus having AzlB as a potential regulator is biologically sound. Another example is motif-96, where the modules are enriched for phage related functions. This is in agreement with the known role of the assigned regulator, Xpf, as controlling the expression from the PBSX prophage late operon (McDonnell & McConnell, 1994).

Table 4.3 Regulons predicted by our integrative analysis (see chapter 5). Predicted motif: motif name; Logo: the corresponding motif logos; Functional Enrichment: the functional enrichment of the modules containing the motif, Module: the module number to which the *de novo* motif was assigned; Candidate TFs: the candidate TFs assigned to the modules based on the principle of 'close genomic neighborhood' and auto-regulation (boxed TFs).

Predicted Motif	Logo	Functional Enrichment	Modules	Candidate TFs
motif_106		Mobility and chemotaxis	67	ykvE rbsR
motif_12		Membrane bioenergetics (electron transport chain and ATP synthase)	1 122 123 124 137 3 46 8	ccpC
motif_14		(Protein synthesis) Ribosomal proteins	63 64	yvdT yqaF ycbG rbsR licT yqaE ywrC yceK
motif_187		(Protein synthesis) Ribosomal proteins	34	ysmB yrxA
motif_2		(Protein synthesis) Ribosomal proteins	39 40	ylbO yotL
motif_260		(Protein synthesis) Ribosomal proteins	121 42	
motif_29		(Protein synthesis) Ribosomal proteins	59	
motif_3		Metabolism of amino acids and related molecules	38	
motif_53		Metabolism of amino acids and related molecules	14 142 27 91 92	arsR ylbO xpf ywgB yopS azlB yvlC yotL
motif_75		Metabolism of amino acids and related molecules	65	ysmB
motif_8		(Protein synthesis) Initiation	112 15	
motif_96		Phage-related functions	48	xpf yotL
motif_98		DNA restriction/modification and repair	30	xpf yopO

4.4 Discussion

As more genome sequences are becoming available, comparative motif detection can serve as a useful independent source of information next to the public experimental data. However, the genome wide application is not trivial for a number of reasons: 1) the phylogenetic step is prone to noise; conserved regions among species are more likely to contain functional sequences but are not restricted to them 2) the comparison of the resulting conserved regions on a genome wide level is tricky 3) clustering methods are usually transitive which results in clustering non-similar regions together.

Therefore, in order to reduce the number of false positives, stringent conditions and filtering steps are needed. For example, we needed to use a very low similarity score (p-value < 0.0001), and clustering was set to deliver only tight clusters, thus many blocks did not cluster and were dropped from the analysis. This could explain the low overlap with the previously defined motifs. In general, this is true for most genome wide phylogenetic footprinting applications.

The problems associated with benchmarking phylogenetic footprinting methods are those of motif detection methods in general. One such category of problems is linked with motif representations. As discussed in chapter 2, the currently used motif representations are highly dependent on the number of contributing sequences, such that both the consensus sequence and the PWM of the motif will differ with changing the constituting sequences. Thus, a systematic comparison of our motifs with the known motifs was difficult to achieve.

The fact that DISTILLER did not validate all of the predicted motifs is partially explained by the limitation of DISTILLER to the available microarrays, such that if some motifs are 'active' in a condition not represented in the microarray compendium, it will not be associated with a module. This is witnessed by the fact that the algorithm missed many of the known motifs as well. Nonetheless, we do not expect that all of the predictions would be functional motifs and some of them are likely false positives.

Chapter 5

5 Inferring the transcriptional network of *Bacillus subtilis*

5.1 Introduction

This chapter describes the research done to infer the transcriptional network of *B. subtilis*. In the introductory chapter, we discussed the significance of understanding the global properties of transcriptional networks as key to an improved perception of the versatile and adaptive lifestyles of bacteria. So far, most studies have been performed on the model organism *E. coli* and have primarily focused on constructing and analyzing the properties of a "static" representation of the transcriptional network: in such representation networks are depicted as directed graphs, in which nodes stand for the regulators and their respective target genes, while the edges indicate the interactions between them. Global features, such as the networks' structural topology, the presence of hubs and network motifs have been largely analyzed (Balaji, Babu & Aravind, 2007; Ernst *et al.*, 2008; Ma, Buer & Zeng, 2004; Martínez-Antonio & Collado-Vides, 2003; Shen-Orr, Milo, Mangan & Alon, 2002; Yu & Gerstein, 2006). Few studies, however, addressed the "dynamic" network, either with respect to time (Ernst, Vainas, Harbison, Simon & Bar-Joseph, 2007) or with respect to environmental conditions (Lemmens *et al.*, 2008; Luscombe *et al.*, 2004). Indeed, most genes undergo conditional regulation rather than a constitutive one.

In this study, we aimed at expanding the transcriptional network of *B. subtilis* and studying its global properties. To this end, we used a data integration workflow for network construction (Lemmens *et al.*, 2008). The importance of integrating different data types has been discussed in chapter 1. The information on the network structure was derived from motif information, which the predicted motifs discussed in the previous chapter are a part of, while the condition dependence of the network was inferred from a compendium of microarray data. Our analysis resulted in enhancing our understanding of the *B. subtilis* regulatory

network by the expansion of the current network on the one hand, and the analysis of its general properties on the other hand.

5.2 Materials and methods

5.2.1 Constructing a motif compendium

Both the *de novo* motif models described in chapter 4 and the models reported in the database of transcriptional regulation in *Bacillus subtilis* (DBTBS) were used to predict novel instances on a genome-wide scale. We used all 159 *de novo* models and 44 models for 44 regulators from DBTBS (33 models of transcription factors and 11 models for alternative sigma factors). Motif screening was performed using the method of Hertzberg *et al.* (Hertzberg, Zuk, Getz & Domany, 2005). We screened the promoter regions of all 4106 protein-coding genes. The promoter region was defined as the intergenic region, plus 100 bp downstream of the ATG.

5.2.2 Construction of the microarray compendium

A compendium of 231 publicly available microarrays were downloaded from the Stanford Microarray Database (Demeter et al., 2007) and Gene Expression Omnibus (Edgar, Domrachev & Lash, 2002), derived from 5 independent studies and performed on 10 different platforms. An overview of the compendium can be found in Table 5.1. The data were normalized as follows: whenever possible, raw intensities were used as data source. Dual-channel data were loess fitted without background correction. Single-channel data were first normalized per experiment (loess fit against an artificial standard consisting of the overall per gene medians), and then log ratios were created using one of the arrays of the series as a reference. The reference array was selected based on its biological role in the experiment. Variance rescaling of the gene expression profiles was performed to render the magnitudes of expression changes more comparable between genes. Microarrays were assigned to six conditional categories: DNA, heat, peroxide, phosphate, quorum response and sporulation, based on a manual curation of the corresponding literature (either a mutant or a condition related to the assigned conditions was altered in the corresponding microarray experiments).

Table 5.1. Overview of microarray experiments used in the study.

Experiment	No. of arrays	Data type	No. of platforms	Data source	Reference (PMID)
GSE1620	15	cDNA	1	GEO	15383836
GSE2667	39	cDNA	1	GEO	16291680
GSE4350	78	cDNA	1	GEO	12486061
GSE4670	24	cDNA & oligo	2	GEO	16816200
GSE4673	72	cDNA & oligo	5	GEO	16855250
peroxide_8uM_t40	3	cDNA	1	SMD	12486061

5.2.3 Inference of transcriptional regulatory modules

Two matrices were used as input for DISTILLER: The first is an expression matrix containing the original values of the compendium converted to percentile ranks (ranging from zero to one). The second is a motif matrix in a binary format, where motif instances with Hertzberg screening p-values ≤ 0.001, and instances documented in DBTBS (including intra-operon regulation) were set to one. For DISTILLER the following parameter settings were used: minimum number of genes in a module = 4; minimum number of conditions in a module = 30; number of randomizations = 100,000; box p-value threshold = 0.001; bandwidth = 0.00001. Post-processing and seed-extension were performed as described in Lemmens *et al*. (submitted).

5.2.4 Benchmarking of the inferred interactions with DBTBS

For benchmarking, only regulator-target interactions for regulators with a documented motif model in DBTBS were considered. Operons were taken into account as follows: for any known regulator binding site of a gene, all genes in the same operon downstream of that gene were also considered to be regulated by the regulator, so they also counted as "true" interactions. To this end the operon assignment of DBTBS was used. Note that this procedure potentially results in an underestimation of the sensitivity of DISTILLER, since some downstream operon genes are not necessarily co-expressed due to condition dependent termination of operon transcription (information which is not available in DBTBS).

5.2.5 Functional and condition enrichment

The functional categories for all genes in a module were downloaded from DBTBS. Module enrichment for a specific functional category was calculated by means of the hypergeometric distribution. For each regulator, functional enrichment was calculated by combining the p-values of enrichment of the modules they regulate, using Fisher's test (Fisher, 1932). A p-value ≤ 0.05 was considered significant. The same method was used to calculate the condition enrichment using the six categories described before.

5.2.6 Prediction of intra-operonic binding sites

Intra-operonic binding sites were identified by searching for operon genes, (according to their experimentally verified annotation in DBTBS), that were assigned to modules different than those of the immediate preceding genes in the operon. The presence of a gene in a different module than that of the immediate upstream gene in the operon is because the gene has its own (predicted) motif site and its expression profile is different from that of its preceding gene, and thus, it is an indication that the gene is regulated differently. All predictions of intra-operonic sites are listed in Appendix 3.

5.3 Results

5.3.1 Network construction

In this study we constructed a global view of the condition-dependent transcriptional network of *B. subtilis* by integrating expression data with motif information. An overview of the analysis flow we used is given in Figure 5.1.

The expression data used in this study consists of a cross platform expression compendium that was compiled from public data, available in SMD and GEO (see materials and methods). It contains 231 microarrays, derived from 5 independent studies, performed on 10 different platforms and testing about 6 different experimental conditions (conditional categories). Information on the transcriptional interactions between a regulator and its target genes was derived from motif information. Motifs, the short regulator-specific sequences upstream of a gene, give indirect information on the interaction between a regulator (transcription factor or sigma factor) and its target genes. Motif information was obtained from two different sources. At first, we collected all regulators for which the motif was documented in the database of transcriptional regulation in *Bacillus subtilis* (DBTBS (Makita, Nakao, Ogasawara & Nakai, 2004)), which were 44 in total. Experimentally verified binding

sites of these regulators were considered as true interactions and included as such in our motif matrix (see materials and methods). Novel sites for the same regulators were predicted using a genome wide screening with the 44 annotated motifs models.

As the regulators with known motif model constitute only a fraction of the total pool of *B. subtilis* regulators, we relied on the *de novo* motifs we detected in chapter 4. For network reconstruction, we used the data integration framework 'DISTILLER'. DISTILLER uses a strategy based on itemset mining to integrate expression with motif data (Lemmens *et al*. submitted). The algorithm searches for gene sets that are co-expressed under a subset of conditions (modules or bi-clusters) and that share common regulatory interactions. It thus identifies regulons, their corresponding regulators and the conditions under which these regulons are active.

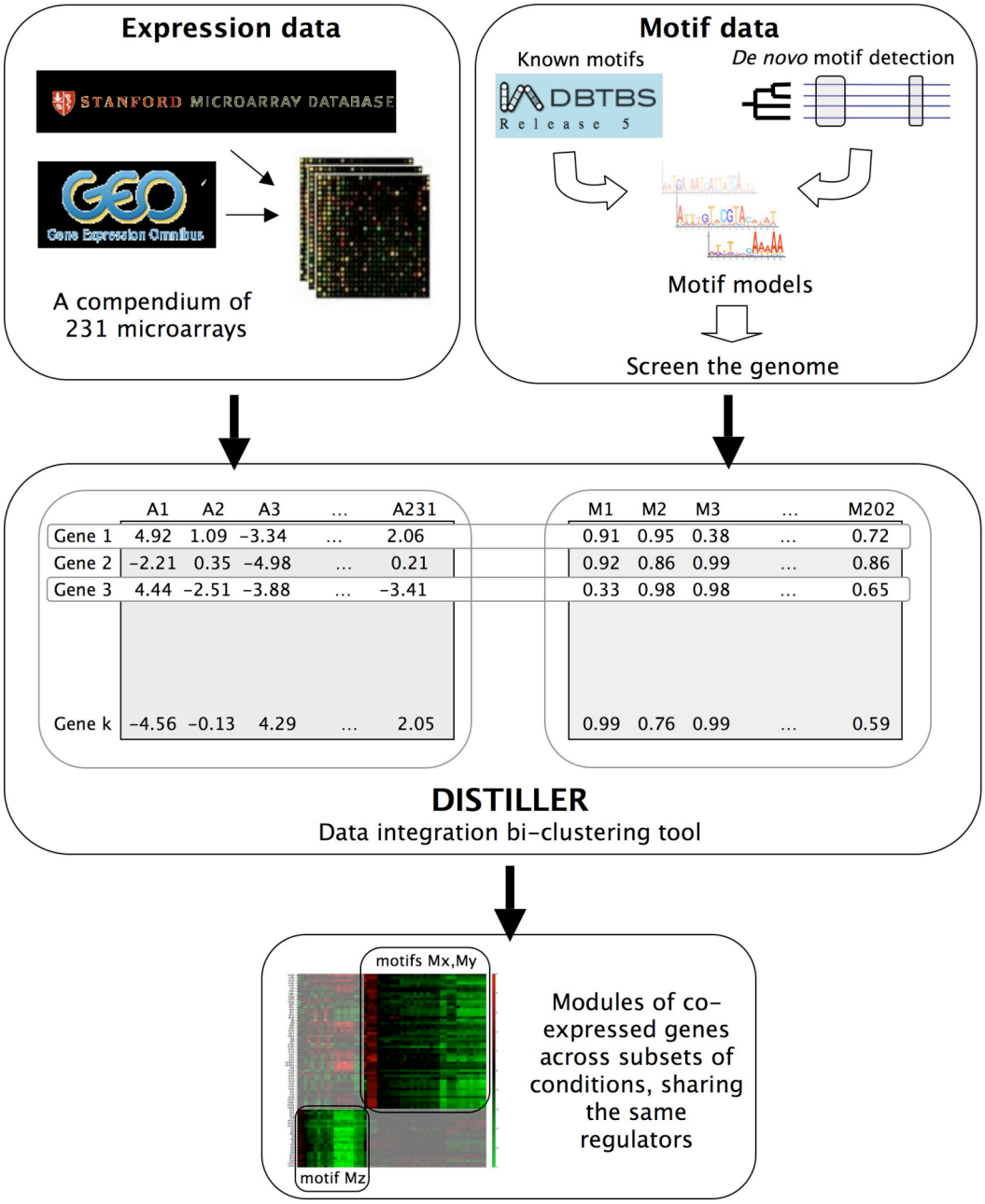


Figure 5.1 Reconstructing the condition-dependent transcriptional network of *B. subtilis*. With DISTILLER information on expression and transcriptional interactions was integrated. Expression data was derived from a cross platform compendium of publicly available microarrays. Interaction information was derived from motif data, and consists of 1) experimentally validated and predicted interactions for all the regulators with a motif model documented in DBTBS, 2) interactions inferred by *de novo* motif detection. The combination of all known and predicted motif-target interactions constitutes the motif data. Applying DISTILLER on the combined motif and expression data resulted in the detection of condition dependent complex regulons, organized in modules.

5.3.2 Network modules and inferred interactions

Combining the expression with the motif information resulted in the detection of 142 modules consisting of 1574 motif-target interactions, for 35 distinct regulatory motifs (22 known and 13 novel) and 1153 genes. Modules range in size between 4-98 genes and 17-87 conditions. Comparing our results with annotated regulator-target interactions in DBTBS showed that we retrieved 44% of the previously characterized interactions, involving 22 (50%) of the documented regulators with known motif model. For the modules corresponding to a known regulator, the module genes were enriched for the same function as the one the regulator is known to be involved in (Figure 5.2). For example, modules of the well-known regulators of sporulation SpoIIID, SigK, and SigE, are mostly enriched for sporulation-related genes. For the other 22 regulators, none of the previously described interactions could be retrieved. The reason for this could be either that these regulators do not meet the minimal requirements needed to obtain their modules, or that they are not triggered in the experimental conditions present in our compendium.

We were able to extend the regulons of the regulators for which DISTILLER found interactions, by an average of 19 targets (see Table 5.2). A detailed list of the interactions predicted per regulator can be found in the supporting Appendix 1. The highest number of predicted novel targets was for the regulators with global functions, such as competence (ComK) and catabolite repression (CcpA, AbrB) (Table 5.1). This is probably partly due to the fact that the conditions under which the microarrays were performed are highly representative for those regulators. On the other hand, it demonstrates that the regulons for these regulators as currently described are still largely incomplete.

Table 5.2 Summary of known and predicted interactions for each of the known regulators assigned to the modules. For all the listed regulators, the size of the known regulon documented by DBTBS, including operon genes is shown. The number of overlapping interactions between this study and the known ones is listed. Also listed are the number of novel interactions predicted by our study and their presumed mode of regulation. Dual regulators are indicated by an *. The cases where we could not predict the mode of regulation are listed as unknown.

Regulator	Size of known regulon	DISTILLER		Predicted mode of regulation for novel interactions		
		Overlap with known interactions	Novel interactions	Positive	Negative	Unknown
AbrB*	87	42	69	0	69	0
CcpA*	130	55	71	0	71	0
CodY	44	8	13	0	13	0
ComA	15	11	0	-	-	-
ComK*	55	34	120	64	0	56
Fur	47	36	20	0	8	12
GerE	33	17	4	0	0	4

LexA	52	31	4	0	4	0
PerR*	19	6	33	0	0	33
PurR	33	21	0	-	-	-
ResD	39	13	12	12	0	0
SigB	101	72	1	1	0	0
SigD	77	61	0	-	-	-
SigE	148	99	21	21	0	0
SigF	40	9	3	3	0	0
SigG	91	32	0	-	-	-
SigH	34	10	6	6	-	-
SigK	100	62	0	-	-	-
SigW	62	37	25	25	0	0
Spo0A*	36	12	3	1	0	2
SpoIIID*	57	31	6	1	3	2
TnrA*	65	5	6	6	0	0

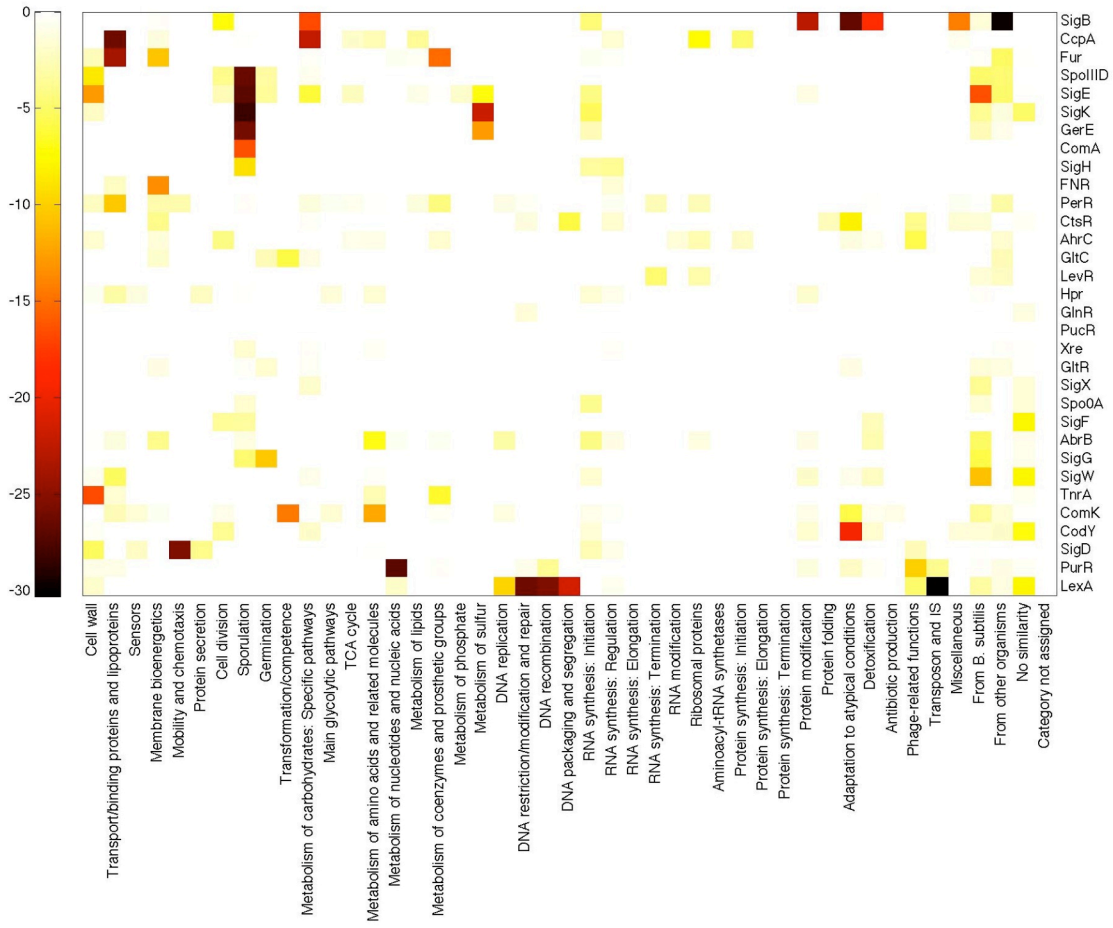


Figure 5.2. Functional enrichment of the retrieved modules. Rows: regulators assigned to at least one module of the output of DISTILLER. Columns: functional categories (according to DBTBS). Shaded boxes indicate functional categories that were statistically enriched in the modules of the corresponding regulator. The color scale represents the extent of enrichment (log p-value) with increased intensity for higher enrichment.

5.3.3 Mode of regulation

Since genes in a module are co-expressed, it is reasonable to assume that they are regulated in a similar mode (activated or repressed) by the assigned regulator. If, for at least three targets in this module, the mode of regulation is known, we could extrapolate this information to the other non-annotated members of the same module. Otherwise, the information was not considered reliable and no extrapolation was made. This inference of the "mode of regulation" is particularly interesting for the dual regulators, i.e. regulators with both an activator or repressor activity (Table 5.2). For some of these dual regulators such as AbrB, CcpA, and ComK we noticed a clear bias towards either a positive or negative mode of regulation (activation or repression, respectively); this might be a consequence of a similar bias present in the known mode of regulation for these regulators (DBTBS). A full list of the novel interactions with their predicted mode of regulation can be found in the supporting Appendix 1.

5.3.4 Condition-dependency of the transcriptional network

The different arrays of the compendium were subdivided into six categories according to the general condition they assessed: DNA, heat, peroxide stress, phosphate starvation, quorum response and sporulation. Microarrays under each category involve either mutants and/or conditions related to the assigned category (see Appendix 2 for the full list of arrays and their conditional categories). We examined whether the arrays of modules regulated by a certain regulator are enriched for particular conditional categories. Enrichment for a certain category implies that the target genes are co-expressed under that conditional category, and indirectly suggests that the regulator is transcriptionally active under that condition. The results indicate that the regulators in our network show a preference of 'activity' towards certain conditions, and that those conditions are generally in agreement with the known role of the regulator (Figure 5.3).

For example, regulators whose modules are enriched for 'DNA' include LexA, the best known regulator of DNA damage-inducible genes (Wojciechowski, Peterson & Love, 1991), and SigD, a regulator of the autolytic response among other functions (Márquez et al., 1990). The induction of autolysis by DNA damage has already been suggested to be the bacterial counterpart of the eukaryotic apoptosis in response to irreparable DNA damage (Bayles, 2003); hence, it is not surprising to see an 'activity' for SigD in the 'DNA' category.

Under conditions of peroxide stress (peroxide category), we notice SigW as the most prominent actor. Indeed, SigW is a regulator of detoxification processes, known to rid the cell from harmful substances (in this case peroxide) (Turner & Helmann, 2000). PerR also exhibits 'activity' in the peroxide category, as would be expected for the repressor of the peroxide regulon (Bsat, Herbig, Casillas-Martinez, Setlow & Helmann, 1998).

The "phosphate" category includes for the most part arrays related to phosphate starvation. Starvation is a known trigger for sporulation (Sonenshein, 2000), and thus it is not a surprise to see the sporulation-specific regulators such as SigE, SigG, SigF and SpoIIID being active in these conditions. ResD, a known target of the phosphate regulator PhoP (Birkey, Liu, Zhang, Duggan & Hulett, 1998) with an activation role in the global regulation of aerobic and anaerobic respiration (Nakano, Zhu, Lacelle, Zhang & Hulett, 2000), is active in this category too.

For the condition 'sporulation', SpoIIID, GerE, SigK, and SigE, all known regulators of sporulation (Piggot, PJ & Hilbert, DW, 2004) appeared to be the main regulators. Other known sporulation-specific regulators such as SigF, SigH, SigG, and Spo0A do not appear to be active in this condition. This is not a surprise, as the conditional category 'sporulation' in our compendium is mainly comprised of experiments performed in the mother cell 5 hours after induction of sporulation, thus excluding the spore-localized regulators SigF and SigG (Eichenberger et al., 2004) and the early-stage sporulation regulators, Spo0A and SigH (Piggot, PJ & Hilbert, DW, 2004).

Some cases of conditional dependency, however, are more perplexing: under quorum response, we were surprised to see a pronounced activity for Fur, the regulator of iron uptake. A recent study reports that some quorum sensing genes in the Gram-positive bacterium *Pseudomonas syringae* are regulated by Fur (Cha, Lee, Oh, Choi & Baik, 2008). A similar situation may be present in *B. subtilis*, which would explain our result.

Examining the conditions under which the *de novo* motifs are 'active', we find that, at least for some of them the conditional categories are in agreement with the functional enrichment of their corresponding modules. For example, motif-96 and motif-98 are both 'active' in the 'DNA' conditional category, while their corresponding modules are enriched for phage-related functions and DNA restriction/modification/repair functions, respectively. However, some cases are less obvious. We notice that three out of the five motifs whose modules are enriched for ribosomal proteins (motif-2, 14, and 187) are 'active' in the phosphate starvation conditional category. It is a known fact in *E. coli,* for instance, that upon phosphate starvation ribosomes are dismantled (Davis, Luger & Tai, 1986) and their

encoding genes undergo a decreased expression (Baek & Lee, 2007). Our results suggest that a similar process to the one in *E. coli* could be taking place in *B. subtilis*.

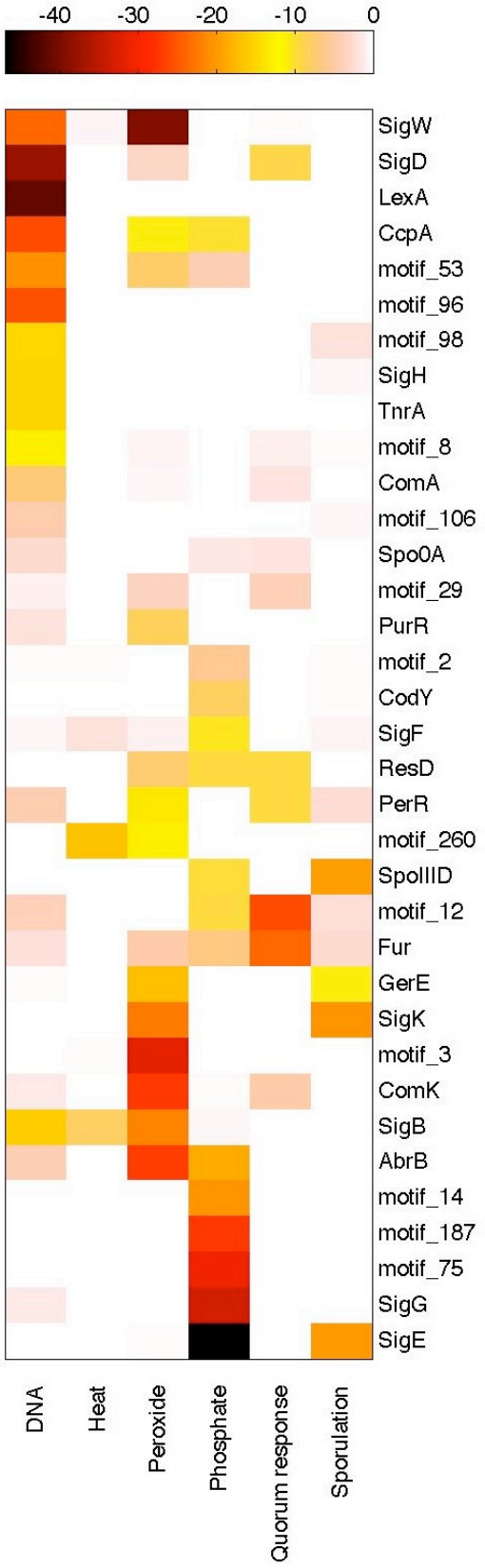


Figure 5.3. Conditional dependency of the network. Rows: regulators or predicted motifs assigned to modules. Columns: condition categories (see supporting table S3 for

the full list of conditional categories of arrays). Shaded boxes indicate conditional categories that were significantly enriched in the modules of the corresponding regulator/motif. The color scale represents the extent of enrichment (log p-value) with increased intensity for higher enrichment.

5.3.5 Prediction of intra-operonic motif sites

In bacteria genes are often organized in single co-regulated transcription units, called operons. It is not uncommon, however, that an operon will have multiple promoters and that a set of genes that acts as a single transcription unit (TU) under one condition is split into various separate units under different conditions. Such split of an operon can occur if the operon contains internal terminators and/or promoter sites (called intra-operonic promoters and characterized by the presence of intra-operonic motif sites; (Laing, Mersinias, Smith & Hubbard, 2006). By combining co-expression with motif information, we can specifically search for novel intra-operonic motifs that give rise to a condition-dependent expression. Therefore, finding an operon gene that has its own motif and that is expressed differently from its preceding genes in the operon (i.e. belongs to a different module than its preceding genes), points towards the presence of a condition-dependent intra-operonic motif site. Accordingly, we retrieved 42 of the 108 intra-operon motif sites reported in DBTBS. In addition, we found evidence for 49 potential new ones (full list in Appendix 3). Figure 5.4 shows an example of the prediction of intra-operonic promoters within the previously characterized operon *hemAXCDBL*. The operon is known to be regulated by a single promoter upstream of *hemA* (Figure 5.4A). We predict the presence of two internal motif sites for the *de novo* predicted motif_187, upstream of *hemX* and *hemL*, respectively (purple boxes in Figure 5.4A). The transcription profile of these two genes is indeed very different from that of the rest of the operon genes (Figure 5.4B).

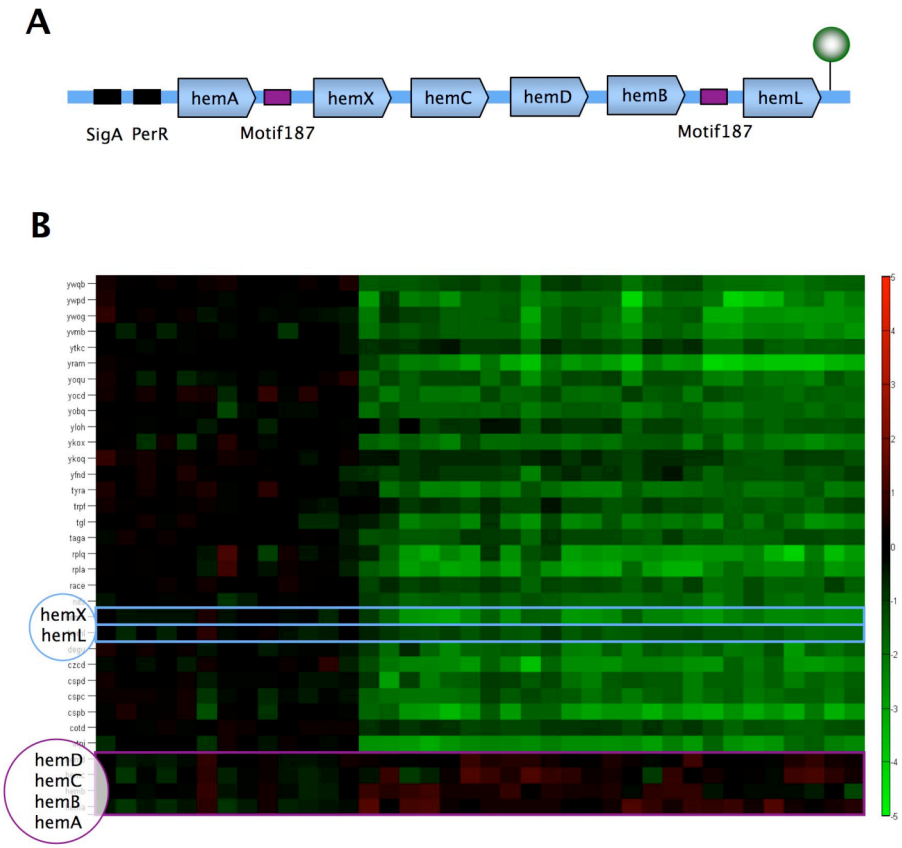


Figure 5.4. Example of predicted intra-operonic motif sites. A: genomic organization of the known operon *hemAXCDBL*. The *hemA, hemX, hemC, hemD, hemB*, and *hemL* genes are represented by arrows, and are known to form an operon (DBTBS) regulated by a single promoter to which SigA and PerR bind (binding sites are shown as black boxes). A terminator (green circle) is present at the end. Two internal motif sites for motif_187 are predicted by our analysis (purple boxes). B: Expression pattern of module 34 genes to which *hemX* and *hemL* belong. The remaining operon genes that are not part of the module are added at the bottom of the figure. All patterns are plotted across the conditions selected for module 34. Rows: genes. Columns: arrays. Module 34 is regulated by motif_187. Genes *hemX* and *hemL* are assigned to this module (enclosed in blue boxes). Their expression profile clearly differs from that of the rest of the operon genes that are not part of the module (enclosed in a purple box at the bottom).

5.3.6 Identification of complex regulons

It is believed that transcriptional complexity is partially mediated by the combined action of regulators: a combination of regulatory binding sites will give rise to more specific expression patterns. To grasp this level of transcriptional complexity, the concept of complex regulons was introduced (Gutiérrez-Ríos et al., 2003) and refers to sets of genes regulated by more than one regulator. DISTILLER has the advantage that it identifies such complex regulons together with their corresponding expression pattern automatically. Given the currently available data, we could identify 6 such regulons, each composed of at least 4 independent transcription units. Five of these complex regulons confirm previously described regulator pairs; SigW-AbrB, SigK-GerE, SigK-SpoIIID, SigE-SpoIIID, and SigW-SigX all have common targets reported in DBTBS. The newly predicted Fur-PerR regulon consists of 30 targets, many of which correspond to previously known Fur targets, but not known to be PerR targets. Fur is a negative regulator of siderophore biosynthesis and of the transcription of ferri-siderophore uptake genes, while PerR is a transcriptional repressor of the peroxide regulon (Bsat et al., 1998). As both regulators have their roles intersecting through the use of iron as a cofactor for antioxidant defense enzymes such as catalase, peroxidase, and superoxide dismutase (Agranoff & Krishna, 1998), it is not surprising to find them as regulators of a common complex regulon.

5.3.7 Identifying global regulators

Global regulators have been defined in different ways in the literature: while some definitions are restricted to having a large regulon size (Babu & Teichmann, 2003; Shen-Orr et al., 2002), others include additional criteria such as the vast number of environmental conditions to which these regulators respond (Gottesman, 1984; Martínez-Antonio & Collado-Vides, 2003) or the number of functional categories to which their target genes belong. Using a combination of criteria we were able to identify the global regulators in our inferred network. The criteria are as follows: 1) regulon size, 2) the range of environmental conditions in which their corresponding modules are enriched (see Condition-dependency of the transcriptional network), 3) the number of functional categories for which their target genes are enriched (see Module networks and inferred interactions). Figure 5.5 contains a color-coded summary of the extent to which each of the regulators that was assigned to a module meets the criteria mentioned above. Four regulators were identified as global based

on a simple scoring system of summing up the scores of the 3 criteria mentioned: AbrB, CcpA, ComK, and SigB have the four highest scores. The known functions of these regulators indeed involve general cellular functions (Table 5.3). We also notice that, with the exception of SigB, the global regulators are dual regulators. This is not a coincidence since in *E. coli* all of the global regulators are also dual regulators (Martínez-Antonio & Collado-Vides, 2003). It seems logical that a regulator that orchestrates the expression of many targets in the cell under different conditions should be equipped with the greatest functional flexibility.

Table 5.3. Functions of the global regulators

Global regulator	Function
AbrB	Regulation of transition state genes
CcpA	Carbon catabolite repression
Comk	Competence
SigB	General stress

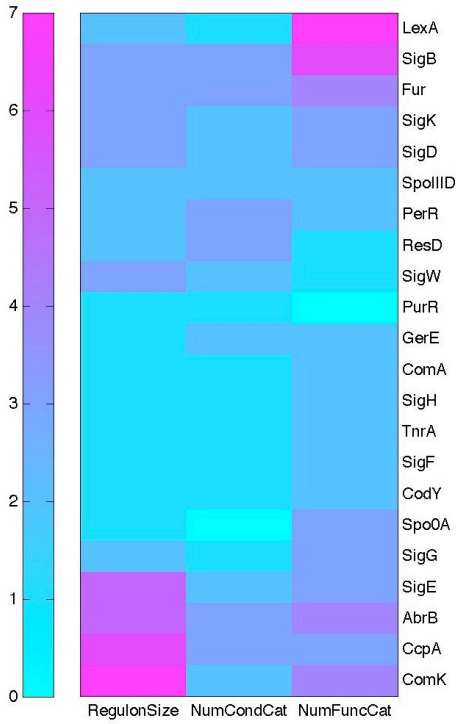


Figure 5.5 Quantitative assessment of different criteria of global regulators for the regulators in this study. Rows: regulators. Columns: criteria for global regulators. RegulonSize = regulon size of a regulator according to our modules, discretized into 7 categories incrementing by 25 targets; NumCondCat = number of conditional categories for which the modules regulated by the regulator were enriched; NumFuncCat = number of functional categories for which the target genes in the modules regulated by the regulator were enriched.

5.3.8 Properties of the regulator interaction network

To have a direct view on the transcriptional information flow from one regulator to another, we reduced the network to a graph in which only the nodes that correspond to regulators are displayed. Of particular interest are the newly inferred interactions between regulators as they unveil potential new links between different transcriptionally regulated pathways (dotted lines in Figure 5.6). For instance, we predict a positive regulation of *sigH* by ComK. Both SigH and ComK are transcriptional regulators known to play a role in cell competence. Specifically, ComK is a major regulator of competence genes, while SigH regulates the transcription of quorum sensing proteins (Hamoen, Venema & Kuipers, 2003), a signaling mechanism known to influence the development of competence in a cell (Tortosa et al., 2001). Another interesting example is the prediction of direct activation of *sigG* expression by SigE. It has been reported that SigE indirectly affects SigG through the transcriptional activation of SpoIIIA, which in turn releases *sigG* from inhibition by an unknown factor (Piggot, PJ & Hilbert, DW, 2004). The direct activation of *sigG* expression by SigE would thus be complementary to this indirect interaction. Worth mentioning is that the genes *sigE* and *sigG* are adjacent on the genome; often times this genomic order entails cross regulation of one gene by the product of the other.

Examining the topology of this network, we see that 50% of the edges belong to only 28% of the nodes. We also found that 50% of the inward edges belong to 27% of the nodes, while 50% of the outward edges belong to 19% of the nodes. The discrepancy between the inward and outward distributions reflects the limited capacity of a gene to receive many regulating signals compared to the capacity of its product to regulate many targets. Furthermore, it was observed in metabolic networks, that there is a positive correlation between the numbers of inward and outward edges per node (Jeong et al., 2000; Schwartz, Cohen, Ben-Avraham, Barabási & Havlin, 2002). In our regulators network no such correlation was found. This could be attributed to the different functionalities of the transcriptional versus the metabolic network: while a metabolic network mainly displays a product/substrate relationship, the transcriptional network reflects how signals are distributed, and this seems not to entail that nodes emitting many (few) signals also receive many (few) signals.

To unravel the hierarchical structure of this network, we followed the approach of Yu *et al*. (Yu & Gerstein, 2006). The authors use a breadth-first-search method in which they first assign all nodes with no outward edges to the first level. Then, they assign all nodes with a direct link to any of the first level nodes to the second level. Out of the remaining nodes, all the nodes with a direct link to the second level nodes are assigned to the third level and so on.

Applying their strategy, we found the network to have a hierarchical structure consisting of 6 levels (Table 5.4). In a hierarchical structure subjects in the higher levels possess bigger 'authority' or effect. In this regulatory network, as was observed with regulatory networks of other organisms (Yu & Gerstein, 2006), we notice that regulating a big number of targets does not automatically place the regulator at the higher hierarchy. For example, regulators with some of biggest known regulons: SigE, ComK, CCpA and AbrB are positioned on only the second level of the hierarchy. Seemingly, the 'big effectors' are not necessarily the ones with big regulons.

The above-mentioned structure, however, excludes all regulators that regulate themselves or that are not transcriptionally regulated at all. The latter ones are thought to be primarily involved in 'reflex' processes, in which they, in order to respond 'quickly' to stimuli, are regulated by protein-protein interactions rather than at the transcriptional level (Yu & Gerstein, 2006).

Table 5.4 Regulators as assigned to the different levels of the hierarchical network structure by using the method of Yu et al. Six levels have been found.

Level	Regulators
6	SenS
5	Hpr PhoP
4	ResD SigF SigH SinR SpoVT
3	FNR Med Rok SigG Spo0A YvrH
2	AbrB ArfM CcpA CcpB CodY ComK DegU Mta PerR PurR SacT SigB SigE SigK SigW SigX SpoIIID TenI TnrA YrzC
1	AcoR AdaA AnsR AraR ArsR AzlB BmrR CcpC CggR CtsR CssR DnaA ExuR FruR Fur GabR GerE GlnR GltC GltR GntR GutR HrcA HutP IolR KdgR KipR LexA LmrA LrpC LytR PyrR RocR SacY SigD SigI SigL SigM SigY TenA TreR Xre YfhP YfiA YlbO YrhI YrxA YtlI YtrA YvbA YwfK

All other TFs not in the tree, i.e. TFs not regulated by, nor regulating any other TFs are thought to be primarily involved in 'reflex' processes, in which they are regulated by protein-protein interaction and respond 'quickly' to stimulations (Yu *et al.*). We found 58 (49%) such regulators in *B. subtilis* compared to 30 (20%) found in *E. coli*. Again, the discrepancy can be attributed to the incompleteness of regulatory information for *B. subtilis*. A good example for a reflex response is metalloregulation by Zur protein. Zur senses the concentration of zinc inside the cell, and in response regulates the expression of zinc-uptake operons, maintaining zinc homeostasis inside the cell.

The second group of regulators is assumed to be involved in the slower 'cogitation' processes, in which the cell needs to 'make decisions'. Sporulation is a multi-step procedure

controlled by TFs that fall into this group. The decision for sporulation is probably made at the level of SigH or higher. SigH in turn activates the level-3 regulator, Spo0A. Yet the commitment to sporulation takes place only upon phosphorylation of Spo0A. Once this has happened, there is no turning back. Spo0A in turn activates level-2 regulators such as SigE that trigger the actual execution process. SigE regulates a number of genes involved in the sporulation process, such as SpoIID, SpoIIM and SpoIIP, responsible for membrane migration during spore-engulfment. This is a clear example of the multi-layer decision-execution process in the cell.

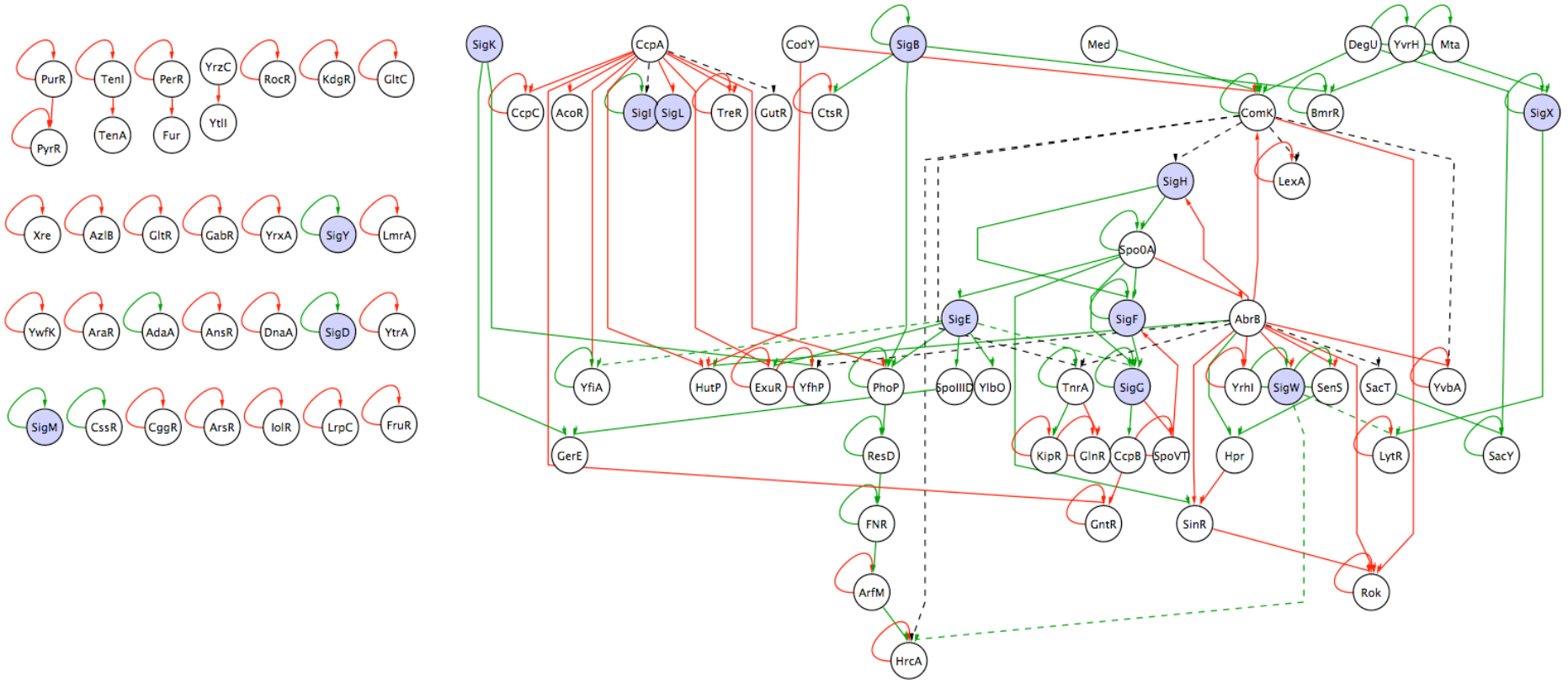


Figure 5.6 The transcriptional network from the regulator point of view. Shown are 85 nodes, representing all regulators for which there is evidence (from this work and from DBTBS) that they regulate other regulators, including auto-regulation. Red edges: negative regulation. Green edges: positive regulation. Black edges: unknown mode of regulation. Solid edges: experimentally confirmed (DBTBS). Dotted edges: predicted in this work. Shaded nodes: sigma factors. White nodes: transcription factors.

5.4 Discussion

In contrast to the *E. coli* and yeast transcriptional networks, few studies exist on the reconstruction of the *B. subtilis* network. In general, these studies focused on a partial reconstruction such as the structure of the sigma factor regulons (De Hoon *et al.* 2004) and the sporulation initiation genetic network (48). Here, we collected different types of data to build a network that would be as inclusive as possible. Using the data integration workflow -DISTILLER- and different data sources as input, we were able to retrieve many of the previously characterized interactions for which the triggering conditions were available in our data set, and confidently predicted 417 new targets for regulators with a known motif model. For many of these novel targets, we could predict their mode of regulation. In addition, several new regulons for yet uncharacterized regulators were predicted based on the output of DISTILLER and *de novo* motif detection by comparative genomics. Our updated network is condition-dependant, providing an insight into which regulatory interactions are ensued under which environmental conditions. This provides a novel view on the network that is not offered by common network construction methods that rely on either the analysis of gene expression for a single condition, or on the examination of the motif data alone. The dependency of the network on conditions was strongly demonstrated through the condition enrichment analysis we performed for each regulator independently; none of the regulators displayed 'activity' across all the conditions, instead they showed preference towards some of them.

The condition-dependency of the network allowed us to also uncover intra-operon regulation. Here, we predicted the presence of 49 potential intra-operonic regulatory sites. Previous studies have predicted operon structures different from those currently listed which are based on experimental evidence. This discrepancy comes as a result of the condition-dependent split of an operon into transcription units. One *B. subtilis* operon prediction study was performed by de Hoon *et al.* in which they identified potential operons based on operon length, intergenic distances, and gene co-expression (49). We viewed their results not as a disagreement with the currently known operon structures, but rather as a prediction of extra regulation within known operons. Based on this, we compared our results for intra-operon regulation with their results and found 13 matches (see Appendix 2). This overlap enforces the validity of our method for predicting intra-operon regulation, and assigns more confidence to those 13 predictions.

In the course of identifying the global regulators of the network, we noticed the role of the alternative sigma factors i.e. non-housekeeping sigma factors. In *B. subtilis* there are 18

known alternative sigma factors, for which 11 have a known motif model. It has been generally accepted that alternative sigma factors provide an effective mechanism for the simultaneous regulation of the expression of large numbers of genes (50,51). However, with the exception of SigB, the 7 alternative sigma factors for which a module was assigned did not fit the former description, nor did they exist on the higher end of the scale of global regulators. The question of the role of regulation by an alternative sigma factor versus that by a transcription factor remains to be addressed.

Our study provides a reliable update of the *B. subtilis* regulatory network and provides insights into its structural properties. Future publication of novel data will certainly enhance the deduction ability of the method and allow for a more complete description of the network of *B. subtilis*.

Chapter 6

6 Conclusions and perspectives

6.1 Conclusions

This thesis has dealt with informatics approaches to solve biological problems. The focus of the research was the regulation of transcription in the cell. To this end, we employed *de novo* motif detection tools to unravel regulator-gene interactions for two bacterial species: *Salmonella* and *Bacillus*. Therefore, the thesis can be roughly divided into two parts examining the two species.

In the first part, a ChIP-chip assay was performed to detect genes bound by HilA, the invasion key regulator in *Salmonella* Typhimurium. The assay revealed a relatively large number of genes bound by HilA, most of which are thought to be false positives. Therefore, the it was complemented with a genome expression profiling of a *hilA* mutant versus a wild type strain, and differentially expressed genes were identified. The intersection of positive hits from the binding and expression assays was extracted, thus delineating genes with a likely functional HilA binding site. Novel motifs were detected in the promoter regions of this list by MDscan, an algorithm that uses ranked lists of sequences to detect motifs. In order to filter out 'false' motifs, the expression data were used as a response factor to gene-motif matching scores in a linear regression analysis, with the algorithm MotifRegressor. An 11-bp long motif was found to best fit the expression profile, and it overlapped with the previously described HilA box. The motif was detected in 9 genes, out of which 3 are known targets of HilA, predicting 4 novel ones.

The second part of the thesis was concerned with reconstructing the transcriptional network of *B. subtilis*. Initially, novel motifs were predicted genome wide, on the basis of comparative genomics and co-regulation. Orthologous genes across four *Bacillus* species were used to perform phylogenetic footprinting. The resulting conserved sequences were compared and clustered, so as to group together all potential binding sites of each TF; co-regulated genes should share the same motif and are expected to fall in the same cluster.

Potential motifs were finally extracted from the clustered sequences of *B. subtilis*, one motif per cluster, making a sum of 159 motifs.

Subsequently, the predicted novel motifs together with the motifs of known regulators documented in DBTBS (44 motifs) were used to screen the entire genome to identify additional binding sites. The outcome was put into a binary matrix where known and predicted binding sites per gene per motif where given a value of one. The motif information was complemented with expression profiling derived from a compendium of 132 public microarrays, and the two data types were used as input for DISTILLER. The latter algorithm employs an itemset mining strategy to identify sets of genes that share the same motifs and are co-expressed in a subset of conditions, i.e. transcriptional modules. 142 modules were inferred, predicting 417 new regulatory interactions for known regulators and 453 interactions for yet uncharacterized regulators. Upon comparison of the inferred interactions with the documented ones in DBTBS, we found a 44% overlap for the regulators used in the analysis. Further validation of the method came about by the finding that the module target genes of a known regulator were enriched for the same function as that known for the regulator.

13 of our *de novo* motifs corresponded to at least one module. However, there is a need to establish a connection between those motifs and regulators. For this purpose, we assigned potential regulators to our predicted motifs based on a known property of bacterial networks that states that a regulator's gene has a distance constraint with its closest binding site on the genome. The pool of regulators from which this assignment took place was a compilation of TFs without a known binding site, and hypothetical TFs derived from a published study.

One of the strong features of DISTILLER is its ability to identify the conditions in which a regulatory interaction takes place. Therefore, the different arrays of the compendium were subdivided into six categories according to the general condition they assessed, and the conditional enrichment of the arrays of each regulator was estimated. The results indicate that the regulators in our network show a preference of 'activity' towards certain conditions, and that those conditions are generally in agreement with the known role of the regulator. In addition, we found evidence for 49 potential condition-dependent intra-operonic motif sites.

The global properties of the network were examined. Global regulators were identified based on their regulon size, the range of environmental conditions they are active in, and the number of functional categories for which their target genes are enriched. Examining the topology of the network revealed a discrepancy between the inward and outward distributions of edges per node, reflecting the limited capacity of a gene to receive many regulating signals compared to the capacity of its product to regulate many targets. Furthermore, in contrast to

the reported positive correlation between the numbers of inward and outward edges per node for metabolic networks, no such correlation was found here.

In brief, the main contribution of this thesis to the area of research is twofold:

- On the biological level it explores for the first time, globally, the transcriptional network of the model organism *B. subtilis*. Although this organism is very well characterized physiologically, biochemically, and genetically, its transcriptional network have only been examined partially. Studying this organism is not only important because of its genetic proximity to other medically important bacilli, e.g. *B. anthracis* the causative agent of anthrax, it also provides a model for a simplified cellular differentiation by undergoing sporulation.
- On the bioinformatics level, this thesis sets a successful example of integrating heterogeneous data sources on the footsteps to the ultimate goal of systems biology. We have exploited sequence information, phylogenetic relationships, known transcriptional interactions, and expression data in one go towards the inference of the regulatory network. Remember that the most important aspect of systems biology is data integration.

6.2 Perspectives

6.2.1 Transcriptional networks involve more complexity than currently accounted for

Modeling of the transcriptional network into a graph of nodes and edges has helped greatly in identifying many of its topological features. However, this representation is far too simplified. To begin with, in the current representation, the nodes signify two different entities at the same time: genes and gene products (TFs). It is well known that there are several checkpoints between the gene and its expression as a fully active protein (regulator). Therefore, the network in fact consists of two layers instead of one: the layer of genes and the layer of functional TFs. Inferring the connections between the two layers requires the assessment of the relationship between the transcriptional regulation of TF's gene (approximated by the levels of its mRNA) and its final form as a functional regulatory protein, in certain conditions. This leads us to the discussion of the condition-dependency of the network addressed below.

Most of the studies view the network as a static structure, where all the edges are present all the time. It is fairly known that only a fraction of the interactions are constitutive,

while the majority of them are temporal, taking place in certain conditions and certain times of a biological process. This dynamic property of the network was demonstrated in this work through the identification of condition 'preferences' of the regulators activity. Also, the study of Luscombe *et al.* (Luscombe et al., 2004) examined the dynamic network on a big scale in yeast, where a wealth of expression data in diverse conditions is available for scientists. The study shows a clear rewiring of the network in different conditions. The same is true for temporal cellular processes such as the cell cycle and sporulation (Ernst, Vainas, Harbison, Simon & Bar-Joseph, 2007). Failure to include the dynamic nature of the network would lead to faulty conclusions.

In addition, the transcription regulation is not a binary function but a continuous one. Regulation of a gene by a certain regulator is not either 'on' or 'off'; there exists the concepts of 'leaky', 'weak', 'tight', and 'strong' promoters, reflecting the gradient of a promoter's regulation. For instance, viral promoters have been typically used in gene therapy because they are known as 'strong' and would guarantee an abundant expression of the implanted gene. In fact, this quantitative regulation is the proposed explanation for the presence of motif degeneracy. Thus, there is a need to model quantitative response of transcription in the network.

Finally, only recently have some studies addresses the 'directed' nature of this network in their models. This is probably due to the lack of proper models in graph theory research to address this issue. Here, we demonstrated that the network topology might be different for the inward and outward edges.

The transcriptional network is a real network that shares a lot of features with other existing real networks such as the WWW and social networks (Barabási & Oltvai, 2004). Advances made on studying dissecting other networks can provide insights into this dynamics of this network, and thus should be taken into account.

6.3 Bridging the gap between biologists and bioinformaticians

The field of bioinformatics existed initially to seal the gap between biology and computer technology. With the advancement of this field, it seems that a gap is opening between biologists and bioinformaticians. The problem for the biologist can be phrased in the following questions:

- Are there any tools out there to solve my problem?
- Which tool is better?

- What is meant by this jargon?
- Which parameters have an influence on my results and what do they mean anyway?

The causes for these problems seem to lie to a large extent within the field of bioinformatics. A relatively young field, it is expected to have not yet arranged its own house, and the requirements of systems biology will put more pressure on it, taking it further into disorder if the proper steps are not taken. The false biological assumptions that some of the present bioinformatics tools are based on demonstrate the poor contact between biologists and bioinformaticians. The presence of many tools to solve the same problem is in many cases caused by bioinformaticians embarking on developing new tools without fully exploring the potentials of existing ones. On the other side of the fence, the biologists seem to lack the technical skills to execute bioinformatics software with the required flexibility. So which party is going to take the step to close the gab?

Here, we suggest a number of solutions to the problem:

- The use of consistent bioinformatics jargon.
- Close collaboration between bioinformaticians and biologists. Bioinformatics cannot stand on its own as a field, at least not yet. This collaboration would have a threefold effect on the bioinformatician: an accurate identification of the biological problem, a solution based on sound biological assumptions, and a proper interpretation of the results. To the biologist, the close contact would inform him of the possibilities that bioinformatics can provide to solve his problems.
- Introduction of user-friendly software with good explanation of the effect of the software parameters on the output. It is not enough that the concepts of the parameters are explained; the user needs to know what effect they have on his analysis.
- Conducting thorough tool comparison studies to evaluate the strengths and weaknesses of each tool. Such kinds of studies have often times revealed the complementarity of the tools rather than the absolute superiority of one tool over the others.
- A proper bioinformatics training of biologists integrated early on in their education. However, biologists should not be overloaded with computer education, and maintain their focus on biology, in order to develop a biologist's intuition.

References

Agranoff, D. D., & Krishna, S. (1998). Metal ion homeostasis and intracellular parasitism. *Molecular Microbiology*, *28*(3), 403-12.

Babu, M. M., & Teichmann, S. A. (2003). Evolution of transcription factors and the gene regulatory network in escherichia coli. *Nucleic Acids Research*, *31*(4), 1234.

Baek, J. H., & Lee, S. Y. (2007). Transcriptome analysis of phosphate starvation response in escherichia coli. *Journal of Microbiology and Biotechnology*, *17*(2), 244-52.

Bailey, T. L., & Elkan, C. (1995). The value of prior knowledge in discovering motifs with MEME. In *Proc int conf intell syst mol biol*.

Balaji, S., Babu, M. M., & Aravind, L. (2007). Interplay between network structures, regulatory modes and sensing mechanisms of transcription factors in the transcriptional regulatory network of E. Coli. *Journal of Molecular Biology*, *372*(4), 1108-22.

Barabási, A. L., & Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nature Reviews. Genetics*, *5*(2), 101-13.

Barash, Y., Elidan, G., Friedman, N., & Kaplan, T. (2003). Modeling dependencies in protein-dna binding sites. In *Proceedings of the seventh annual international conference on research in computational molecular biology*.

Bayles, K. W. (2003). Are the molecular strategies that control apoptosis conserved in bacteria?. *Trends in Microbiology*, *11*(7), 306-11.

Belitsky, B. R., Gustafsson, M. C., Sonenshein, A. L., & Von Wachenfeldt, C. (1997). An lrp-like gene of bacillus subtilis involved in branched-chain amino acid transport. *Journal of Bacteriology*, *179*(17), 5448-57.

Benos, P. V., Bulyk, M. L., & Stormo, G. D. (2002). Additivity in protein-dna interactions: How good an approximation is it?. *Nucleic Acids Research*, *30*(20), 4442-51.

Berezikov, E., Guryev, V., & Cuppen, E. (2005). CONREAL web server: Identification and visualization of conserved transcription factor binding sites. *Nucleic Acids Research*, *33*(Web Server issue), W447-50.

Berezikov, E., Guryev, V., Plasterk, R. H., & Cuppen, E. (2004). CONREAL: Conserved regulatory elements anchored alignment algorithm for identification of transcription factor binding sites by phylogenetic footprinting. *Genome Research*, *14*(1), 170-8.

Berger, M. F., Philippakis, A. A., Qureshi, A. M., He, F. S., Estep, P. W., & Bulyk, M. L. (2006). Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nature Biotechnology*, *24*(11), 1429-35.

Birkey, S. M., Liu, W., Zhang, X., Duggan, M. F., & Hulett, F. M. (1998). Pho signal transduction network reveals direct transcriptional regulation of one two-component system

by another two-component regulator: Bacillus subtilis phop directly regulates production of resd. *Molecular Microbiology*, *30*(5), 943-53.

Blanchette, M., & Tompa, M. (2003). Footprinter: A program designed for phylogenetic footprinting. *Nucleic Acids Research*, *31*(13), 3840.

Blanchette, M., Schwikowski, B., & Tompa, M. (2000). An exact algorithm to identify motifs in orthologous sequences from multiple species. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, *8*, 37-45.

Boffelli, D., McAuliffe, J., Ovcharenko, D., Lewis, K. D., Ovcharenko, I., Pachter, L., et al. (2003). Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science (New York, N.Y.)*, *299*(5611), 1391-4.

Boor, K. J. (2006). Bacterial stress responses: What doesn't kill them can make then stronger. *Plos Biology*, *4*(1), e23.

Brilli, M., Fani, R., & Lió, P. (2007). Motifscorer: Using a compendium of microarrays to identify regulatory motifs. *Bioinformatics (Oxford, England)*, *23*(4), 493-5.

Bsat, N., Herbig, A., Casillas-Martinez, L., Setlow, P., & Helmann, J. D. (1998). Bacillus subtilis contains multiple fur homologues: Identification of the iron uptake (fur) and peroxide regulon (perr) repressors. *Molecular Microbiology*, *29*(1), 189-98.

Buhler, J., & Tompa, M. (2002). Finding motifs using random projections. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, *9*(2), 225-42.

Bulyk, M. L., Johnson, P. L., & Church, G. M. (2002). Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Research*, *30*(5), 1255-61.

Burkovski, A. (2002). Bacterial amino acid transport proteins: Occurrence, functions, and significance for biotechnological applications. *Applied Microbiology and Biotechnology*, *58*(3), 265-274.

Bussemaker, H. J., Li, H., & Siggia, E. D. (2000). Building a dictionary for genomes: Identification of presumptive regulatory sites by statistical analysis. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(18), 10096-100.

Bussemaker, H. J., Li, H., & Siggia, E. D. (2001). Regulatory element detection using correlation with expression. *Nature Genetics*, *27*(2), 167-71.

Carvalho, A. M., Freitas, A. T., Oliveira, A. L., & Sagot, M. F. (2006). An efficient algorithm for the identification of structured motifs in DNA promoter sequences. *IEEE/ACM Transactions on Computational Biology and Bioinformatics / IEEE, ACM*, *3*(2), 126-40.

Cha, J. Y., Lee, J. S., Oh, J. I., Choi, J. W., & Baik, H. S. (2008). Functional analysis of the role of fur in the virulence of pseudomonas syringae pv. Tabaci 11528: Fur controls expression of genes involved in quorum-sensing. *Biochemical and Biophysical Research Communications*, *366*(2), 281-7.

Chen, G., Hata, N., & Zhang, M. Q. (2004). Transcription factor binding element detection using functional clustering of mutant expression data. *Nucleic Acids Research*, *32*(8), 2362-71.

Chen, X., Guo, L., Fan, Z., & Jiang, T. (2008). W-Alignace: An improved gibbs sampling algorithm based on more accurate position weight matrices learned from sequence and gene expression/chip-chip data. *Bioinformatics (Oxford, England)*, *24*(9), 1121-8.

Chen, X., Hughes, T. R., & Morris, Q. (2007). Rankmotif++: A motif-search algorithm that accounts for relative ranks of k-mers in binding transcription factors. *Bioinformatics (Oxford, England)*, *23*(13), i72-9.

Cliften, P., Sudarsanam, P., Desikan, A., Fulton, L., Fulton, B., Majors, J., et al. (2003). Finding functional features in saccharomyces genomes by phylogenetic footprinting. *Science (New York, N.Y.)*, *301*(5629), 71-6.

Conlon, E. M., Liu, X. S., Lieb, J. D., & Liu, J. S. (2003). Integrating regulatory motif discovery and genome-wide expression analysis. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(6), 3339-44.

D'haeseleer, P. (2006). How does DNA sequence motif discovery work?. *Nature Biotechnology*, *24*, 959-961.

Dai, X., He, J., & Zhao, X. (2007). A new systematic computational approach to predicting target genes of transcription factors. *Nucleic Acids Research*, *35*(13), 4433-40.

Das, M. K., & Dai, H. K. (2007). A survey of DNA motif finding algorithms. *BMC Bioinformatics*, *8 Suppl 7*, S21.

Davis, B. D., Luger, S. M., & Tai, P. C. (1986). Role of ribosome degradation in the death of starved escherichia coli cells. *Journal of Bacteriology*, *166*(2), 439-45.

Demeter, J., Beauheim, C., Gollub, J., Hernandez-Boussard, T., Jin, H., Maier, D., et al. (2007). The stanford microarray database: Implementation of new analysis tools and open source release of software. *Nucleic Acids Research*, *35*(Database issue), D766.

Eden, E., Lipson, D., Yogev, S., & Yakhini, Z. (2007). Discovering motifs in ranked lists of DNA sequences. *Plos Computational Biology*, *3*(3), e39.

Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, *30*(1), 207.

Eichenberger, P., Fujita, M., Jensen, S. T., Conlon, E. M., Rudner, D. Z., Wang, S. T., et al. (2004). The program of gene transcription for a single differentiating cell type during sporulation in bacillus subtilis. *Plos Biology*, *2*(10), e328.

Elnitski, L., Hardison, R. C., Li, J., Yang, S., Kolbe, D., Eswara, P., et al. (2003). Distinguishing regulatory DNA from neutral sites. *Genome Research*, *13*(1), 64-72.

Ernst, J., Beg, Q. K., Kay, K. A., Balázsi, G., Oltvai, Z. N., & Bar-Joseph, Z. (2008). A semi-supervised method for predicting transcription factor-gene interactions in escherichia coli. *Plos Computational Biology*, *4*(3), e1000044.

Ernst, J., Vainas, O., Harbison, C. T., Simon, I., & Bar-Joseph, Z. (2007). Reconstructing dynamic regulatory maps. *Molecular Systems Biology*, *3*, 74.

Eskin, E., & Pevzner, P. A. (2002). Finding composite regulatory patterns in DNA sequences. *Bioinformatics (Oxford, England)*, *18 Suppl 1*, S354-63.

Fisher, R. A. (1932). Statistical methods for research workers.

Fratkin, E., Naughton, B. T., Brutlag, D. L., & Batzoglou, S. (2006). Motifcut: Regulatory motifs finding with maximum density subgraphs. *Bioinformatics (Oxford, England)*, *22*(14), e150-7.

Gelfand, M. S., Koonin, E. V., & Mironov, A. A. (2000). Prediction of transcription regulatory sites in archaea by a comparative genomic approach. *Nucleic Acids Research*, *28*(3), 695-705.

Gottesman, S. (1984). Bacterial regulation: Global regulatory networks. *Annual Reviews in Genetics*, *18*(1), 415-441.

GuhaThakurta, D. (2006). Computational identification of transcriptional regulatory elements in DNA sequence. *Nucleic Acids Research*, *34*(12), 3585.

Gupta, M., & Liu, J. S. (2005). De novo cis-regulatory module elicitation for eukaryotic genomes. *Proceedings of the National Academy of Sciences*, *102*(20), 7079-7084.

Gutiérrez-Ríos, R. M., Rosenblueth, D. A., Loza, J. A., Huerta, A. M., Glasner, J. D., Blattner, F. R., et al. (2003). Regulatory network of escherichia coli: Consistency between literature knowledge and microarray profiles. *Genome Research*, *13*(11), 2435-43.

Hamoen, L. W., Venema, G., & Kuipers, O. P. (2003). Controlling competence in bacillus subtilis: Shared use of regulators. *Microbiology (Reading, England)*, *149*(Pt 1), 9-17.

Hertz, G. Z., & Stormo, G. D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics (Oxford, England)*, *15*(7-8), 563-77.

Hertzberg, L., Zuk, O., Getz, G., & Domany, E. (2005). Finding motifs in promoter regions. *Journal of Computational Biology*, *12*(3), 314-330.

De Hoon, M. J., Imoto, S., Kobayashi, K., Ogasawara, N., & Miyano, S. (2004). Predicting the operon structure of bacillus subtilis using operon length, intergene distance, and gene expression information. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 276-87.

Hu, J., Li, B., & Kihara, D. (2005). Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Research*, *33*(15), 4899.

Jensen, S. T., Shen, L., & Liu, J. S. (2005). Combining phylogenetic motif discovery and motif clustering to predict co-regulated genes. *Bioinformatics (Oxford, England)*, *21*(20), 3832-9.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., Barabasi, A. L., & others, O. (2000). The large-scale organization of metabolic networks. *Nature*, *407*(6804), 651-653.

de Jong, H., Geiselmann, J., Hernandez, C., & Page, M. (2003). Genetic network analyzer: Qualitative simulation of genetic regulatory networks. *Bioinformatics (Oxford, England)*, *19*(3), 336-44.

Joshi, A., Van de Peer, Y., & Michoel, T. (2008). Analysis of a gibbs sampler method for model-based clustering of gene expression data. *Bioinformatics (Oxford, England)*, *24*(2), 176.

Kaplan, T., Friedman, N., & Margalit, H. (2005). Ab initio prediction of transcription factor targets using structural knowledge. *Plos Computational Biology*, *1*(1), e1.

Kazmierczak, M. J., Wiedmann, M., & Boor, K. J. (2005). Alternative sigma factors and their roles in bacterial virulence. *Microbiology and Molecular Biology Reviews*, *69*(4), 527-543.

De Keersmaecker, S. C., Marchal, K., Verhoeven, T. L., Engelen, K., Vanderleyden, J., & Detweiler, C. S. (2005). Microarray analysis and motif detection reveal new targets of the salmonella enterica serovar typhimurium hila regulatory protein, including hila itself. *Journal of Bacteriology*, *187*(13), 4381-91.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., & Lander, E. S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, *423*(6937), 241-54.

Kingsford, C., Zaslavsky, E., & Singh, M. (2006). A compact mathematical programming formulation for DNA motif finding. *LECTURE NOTES IN COMPUTER SCIENCE*, *4009*, 233.

Kolbe, D., Taylor, J., Elnitski, L., Eswara, P., Li, J., Miller, W., et al. (2004). Regulatory potential scores from genome-wide three-way alignments of human, mouse, and rat. *Genome Research*, *14*(4), 700-7.

Kummerfeld, S. K., & Teichmann, S. A. (2006). DBD: A transcription factor prediction database. *Nucleic Acids Research*, *34*(Database issue), D74-81.

Laing, E., Mersinias, V., Smith, C. P., & Hubbard, S. J. (2006). Analysis of gene expression in operons of streptomyces coelicolor. *Genome Biology*, *7*(6), R46.

Lawrence, C. E., & Reilly, A. A. (1990). An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins*, *7*(1), 41-51.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., & Wootton, J. C. (1993). Detecting subtle sequence signals: A gibbs sampling strategy for multiple alignment. *Science*, *262*(5131), 208-14.

Lemmens, K., De Bie, T., Dhollander, T., De Keersmaecker, S., Thijs, I., Schoofs, G., et al. (2009). DISTILLER: A data integration framework to reveal condition dependency of complex regulons in escherichia coli. *Genome Biology. Submitted* .

Lemmens, K., Dhollander, T., De Bie, T., Monsieurs, P., De Moor, B., Collado-Vides, J., et al. (2008). The condition-dependent transcriptional network in escherichia coli. *Ann N Y Acad Sci*.

Leung, J. Y., McKenzie, F. E., Uglialoro, A. M., Flores-Villanueva, P. O., Sorkin, B. C., Yunis, E. J., et al. (2000). Identification of phylogenetic footprints in primate tumor necrosis factor-alpha promoters. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(12), 6614-8.

Levy, S., Hannenhalli, S., & Workman, C. (2001). Enrichment of regulatory signals in conserved non-coding genomic sequence. *Bioinformatics (Oxford, England)*, *17*(10), 871-7.

Liu, F. F. M., Tsai, J. J. P., Chen, R. M., Chen, S. N., & Shih, S. H. (2004). FMGA: Finding motifs by genetic algorithm. In *Proc. BIBE.*

Liu, X. S., Brutlag, D. L., & Liu, J. S. (2002). An algorithm for finding protein- DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature Biotechnology*, *20*, 835-839.

Liu, X., Brutlag, D. L., & Liu, J. S. (2001). Bioprospector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. In *Pac. Symp. Biocomput.*

Liu, Z., Guo, J. T., Li, T., & Xu, Y. (2008). Structure-Based prediction of transcription factor binding sites using a protein-dna docking approach. *Proteins*, *72*(4), 1114-24.

Loots, G. G., Locksley, R. M., Blankespoor, C. M., Wang, Z. E., Miller, W., Rubin, E. M., et al. (2000). Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science*, *288*(5463), 136.

Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A., & Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, *431*(7006), 308-12.

Ma, H. W., Buer, J., & Zeng, A. P. (2004). Hierarchical structure and modules in the escherichia coli transcriptional regulatory network revealed by a new top-down approach. *BMC Bioinformatics*, *5*, 199.

MacIsaac, K. D., & Fraenkel, E. (2006). Practical strategies for discovering regulatory DNA sequence motifs. *Plos Computational Biology*, *2*(4), e36.

Makita, Y., Nakao, M., Ogasawara, N., & Nakai, K. (2004). DBTBS: Database of transcriptional regulation in bacillus subtilis and its contribution to comparative genomics. *Nucleic Acids Research*, *32*(Database issue), D75-7.

Mandel-Gutfreund, Y., Baron, A., & Margalit, H. (2001). A structure-based approach for prediction of protein binding sites in gene upstream regions. In *Pac. Symp. Biocomput.*

Manen, J. F., Savolainen, V., & Simon, P. (1994). The atpb and rbcl promoters in plastid dnas of a wide dicot range. *J Mol Evol*, *38*(6), 577-82.

Marsan, L., & Sagot, M. F. (2000). Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. *Journal of Computational Biology*, *7*(3-4), 345-362.

Martínez-Antonio, A., & Collado-Vides, J. (2003). Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr Opin Microbiol*, *6*(5), 482-9.

Márquez, L. M., Helmann, J. D., Ferrari, E., Parker, H. M., Ordal, G. W., & Chamberlin, M. J. (1990). Studies of sigma d-dependent functions in bacillus subtilis. *Journal of Bacteriology*, *172*(6), 3435-43.

McCue, L. A., Thompson, W., Carmack, C. S., Ryan, M. P., Liu, J. S., Derbyshire, V., et al. (2001). Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Research*, *29*(3), 774.

McDonnell, G. E., & McConnell, D. J. (1994). Overproduction, isolation, and dna-binding characteristics of xre, the repressor protein from the bacillus subtilis defective prophage PBSX. *Journal of Bacteriology*, *176*(18), 5831-5834.

McGuire, A. M., Hughes, J. D., & Church, G. M. (2000). Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Research*, *10*(6), 744-57.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., & Alon, U. (2002). Network motifs: Simple building blocks of complex networks. *Science (New York, N.Y.)*, *298*(5594), 824-7.

Monsieurs, P., Thijs, G., Fadda, A. A., De Keersmaecker, S. C., Vanderleyden, J., De Moor, B., et al. (2006). More robust detection of motifs in coexpressed genes by using phylogenetic information. *BMC Bioinformatics*, *7*, 160.

Morozov, A. V., Havranek, J. J., Baker, D., & Siggia, E. D. (2005). Protein-Dna binding specificity predictions with structural models. *Nucleic Acids Research*, *33*(18), 5781.

MOSES, A. M., CHIANG, D. Y., & EISEN, M. B. (2003). PHYLOGENETIC MOTIF DETECTION BY EXPECTATION-MAXIMIZATION ON EVOLUTIONARY MIXTURES. In *Pacific symposium on biocomputing 2004: Proceedings of the pacific symposium.*

Nakano, M. M., Zhu, Y., Lacelle, M., Zhang, X., & Hulett, F. M. (2000). Interaction of resd with regulatory regions of anaerobically induced genes in bacillus subtilis. *Molecular Microbiology*, *37*(5), 1198-207.

Newberg, L. A., Thompson, W. A., Conlan, S., Smith, T. M., McCue, L. A., & Lawrence, C. E. (2007). A phylogenetic gibbs sampler that yields centroid solutions for cis-regulatory site prediction. *Bioinformatics (Oxford, England)*, *23*(14), 1718-27.

O'Flanagan, R. A., Paillard, G., Lavery, R., & Sengupta, A. M. (2005). Non-Additivity in protein-dna binding. *Bioinformatics (Oxford, England)*, *21*(10), 2254-63.

Pavesi, G., Mauri, G., & Pesole, G. (2001). An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics (Oxford, England)*, *17 Suppl 1*, S207-14.

Pavesi, G., Mereghetti, P., Mauri, G., & Pesole, G. (2004). Weeder web: Discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Research*, *32*(Web Server issue), W199-203.

Pavesi, G., Zambelli, F., & Pesole, G. (2007). Weederh: An algorithm for finding conserved regulatory motifs and regions in homologous sequences. *BMC Bioinformatics*, *8*, 46.

Pevzner, P. A., & Sze, S. H. (2000). Combinatorial approaches to finding subtle signals in DNA sequences. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, *8*, 269-78.

Piggot, PJ, & Hilbert, DW (2004). Sporulation of bacillus subtilis. *Current Opinion in Microbiology*, *7*(6), 579-584.

PRAKASH, A., BLANCHETTE, M., SINHA, S., & TOMPA, M. (2003). MOTIF DISCOVERY IN HETEROGENEOUS SEQUENCE DATA. In *Pacific symposium on biocomputing 2004: Proceedings of the pacific symposium.*

Pudimat, R., Schukat-Talamazzini, E. G., & Backofen, R. (2004). Feature based representation and detection of transcription factor binding sites. In *Proceedings of the german conference on bioinformatics*.

Ramsden, A. E., Mota, L. J., Munter, S., Shorte, S. L., & Holden, D. W. (2007). The SPI-2 type III secretion system restricts motility of salmonella-containing vacuoles. *Cellular Microbiology*, *9*(10), 2517-2529.

Reddy, T. E., DeLisi, C., & Shakhnovich, B. E. (2007). Binding site graphs: A new graph theoretical framework for prediction of transcription factor binding sites. *Plos Computational Biology*, *3*(5), e90.

Roth, F. P., Hughes, J. D., Estep, P. W., & Church, G. M. (1998). Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mrna quantitation. *Nature Biotechnology*, *16*(10), 939-45.

Sabatti, C., & Lange, K. (2002). Genomewide motif identification using a dictionary model. *Proceedings of the IEEE*, *90*(11), 1803-1810.

Sabatti, C., Rohlin, L., Lange, K., & Liao, J. C. (2005). Vocabulon: A dictionary model approach for reconstruction and localization of transcription factor binding sites. *Bioinformatics (Oxford, England)*, *21*(7), 922-931.

Sagot, M. F. (1998). Spelling approximate repeated or common motifs using a suffix tree. *LECTURE NOTES IN COMPUTER SCIENCE*, 374-390.

Sandve, G. K., & Drabløs, F. (2006). A survey of motif discovery methods in an integrated framework. *Biology Direct*, *1*, 11.

Sauer, U., Heinemann, M., & Zamboni, N. (2007). Genetics. Getting closer to the whole picture. *Science (New York, N.Y.)*, *316*(5824), 550-1.

Schwartz, N., Cohen, R., Ben-Avraham, D., Barabási, A. L., & Havlin, S. (2002). Percolation in directed scale-free networks. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, *66*(1 Pt 2), 015104.

Segal, E., Shapira, M., Regev, A., Pe er, D., Botstein, D., Koller, D., et al. (2003). Module networks: Identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics*, *34*(2), 166-176.

Shen-Orr, S. S., Milo, R., Mangan, S., & Alon, U. (2002). Network motifs in the transcriptional regulation network of escherichia coli. *Nature Genetics*, *31*(1), 64-68.

Shida, K. (2006). Gibbsst: A gibbs sampling method for motif discovery with enhanced resistance to local optima. *BMC Bioinformatics*, *7*, 486.

Siddharthan, R., Siggia, E. D., & van Nimwegen, E. (2005). Phylogibbs: A gibbs sampling motif finder that incorporates phylogeny. *Plos Computational Biology*, *1*(7), e67.

Sinha, S., & Tompa, M. (2000). A statistical method for finding transcription factor binding sites. In *Proceedings of the eighth international conference on intelligent systems for molecular biology*.

Sinha, S., Blanchette, M., & Tompa, M. (2004). Phyme: A probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics*, *5*, 170.

Sinha, S., Liang, Y., & Siggia, E. (2006). Stubb: A program for discovery and analysis of cis-regulatory modules. *Nucleic Acids Research*, *34*(Web Server issue), W555-9.

Sonenshein, A. L. (2000). Control of sporulation initiation in bacillus subtilis. *Current Opinion in Microbiology*, *3*(6), 561-566.

Sosinsky, A., Honig, B., Mann, R. S., & Califano, A. (2007). Discovering transcriptional regulatory regions in drosophila by a nonalignment method for phylogenetic footprinting. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(15), 6305-10.

Stormo, G. D. (2000). DNA binding sites: Representation and discovery. *Bioinformatics (Oxford, England)*, *16*(1), 16-23.

Stormo, G. D., & Hartzell, G. W. (1989). Identifying protein-binding sites from unaligned DNA fragments. *Proceedings of the National Academy of Sciences of the United States of America*, *86*(4), 1183-7.

Tagle, D. A., Koop, B. F., Goodman, M., Slightom, J. L., Hess, D. L., & Jones, R. T. (1988). Embryonic epsilon and gamma globin genes of a prosimian primate (galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *Journal of Molecular Biology*, *203*(2), 439-55.

Tan, K., McCue, L. A., & Stormo, G. D. (2005). Making connections between novel transcription factors and their DNA motifs. *Genome Research*, *15*(2), 312-20.

Thijs, G., Marchal, K., Lescot, M., Rombauts, S., De Moor, B., Rouzé, P., et al. (2002). A gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, *9*(2), 447-64.

Thijs, I. M., De Keersmaecker, S. C., Fadda, A., Engelen, K., Zhao, H., McClelland, M., et al. (2007). Delineation of the salmonella enterica serovar typhimurium hila regulon through genome-wide location and transcript analysis. *Journal of Bacteriology*, *189*(13), 4587-96.

Thompson, W. A., Newberg, L. A., Conlan, S., McCue, L. A., & Lawrence, C. E. (2007). The gibbs centroid sampler. *Nucleic Acids Research*, *35*(Web Server issue), W232-7.

Tompa, M. (1999). An exact method for finding short motifs in sequences, with application to the ribosome binding site problem. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, 262-71.

Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., et al. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, *23*(1), 137-44.

Tortosa, P., Logsdon, L., Kraigher, B., Itoh, Y., Mandic-Mulec, I., & Dubnau, D. (2001). Specificity and genetic polymorphism of the bacillus competence quorum-sensing system. *Journal of Bacteriology*, *183*(2), 451-60.

Turner, M. S., & Helmann, J. D. (2000). Mutations in multidrug efflux homologs, sugar isomerases, and antimicrobial biosynthesis genes differentially elevate activity of the sigma(X) and sigma(W) factors in bacillus subtilis. *Journal of Bacteriology*, *182*(18), 5202-10.

van Helden, J., André, B., & Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *Journal of Molecular Biology*, *281*(5), 827-42.

Vanet, A., Marsan, L., Labigne, A., & Sagot, M. F. (2000). Inferring regulatory elements from a whole genome. An analysis of helicobacter pylori sigma(80) family of promoter signals. *Journal of Molecular Biology*, *297*(2), 335-53.

Vuillaumier, S., Dixmeras, I., Messaï, H., Lapouméroulie, C., Lallemand, D., Gekas, J., et al. (1997). Cross-Species characterization of the promoter region of the cystic fibrosis transmembrane conductance regulator gene reveals multiple levels of regulation. *The Biochemical Journal*, *327 (Pt 3)*, 651-62.

Wang, G., Yu, T., & Zhang, W. (2005). Wordspy: Identifying transcription factor binding motifs by building a dictionary and learning a grammar. *Nucleic Acids Research*, *33*(Web Server issue), W412-6.

Wang, T., & Stormo, G. D. (2003). Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics (Oxford, England)*, *19*(18), 2369-80.

Wang, T., & Stormo, G. D. (2005). Identifying the conserved network of cis-regulatory sites of a eukaryotic genome. *Proceedings of the National Academy of Sciences*, *102*(48), 17400-17405.

Wasserman, W. W., & Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, *5*(4), 276-287.

Wasserman, W. W., Palumbo, M., Thompson, W., Fickett, J. W., & Lawrence, C. E. (2000). Human-Mouse genome comparisons to locate regulatory sites. *Nature Genetics*, *26*(2), 225-8.

Wei, Z., & Jensen, S. T. (2006). GAME: Detecting cis-regulatory elements using a genetic algorithm. *Bioinformatics (Oxford, England)*, *22*(13), 1577-84.

Wojciechowski, M. F., Peterson, K. R., & Love, P. E. (1991). Regulation of the SOS response in bacillus subtilis: Evidence for a lexa repressor homolog. *Journal of Bacteriology*, *173*(20), 6489-98.

Workman, C. T., & Stormo, G. D. (2000). Ann-Spec: A method for discovering transcription factor binding sites with improved specificity. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 467-78.

Yu, H., & Gerstein, M. (2006). Genomic analysis of the hierarchical structure of regulatory networks. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(40), 14724-31.

Zhang, Z., Schwartz, S., Wagner, L., & Miller, W. (2000). A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology*, *7*(1-2), 203-14.

Zhou, Q., & Liu, J. S. (2004). Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics (Oxford, England)*, *20*(6), 909-16.

Appendix 1

List of predicted novel interactions and their mode of regulation. Each predicted regulator-gene interaction is listed in the corresponding columns. A positive or negative mode of regulation is listed in the column 'Regulation', as well as the module in which this interaction was predicted.

Regulator	Gene	Regulation	Module
AbrB	appB	Negative	12
AbrB	appC	Negative	12
AbrB	appD	Negative	12
AbrB	pspA	Negative	12
AbrB	ydjG	Negative	12
AbrB	ydjI	Negative	12
AbrB	ywoF	Negative	12
AbrB	pksC	Negative	37
AbrB	pksD	Negative	37
AbrB	qcrA	Negative	37
AbrB	qcrB	Negative	37
AbrB	qcrC	Negative	37
AbrB	rapG	Negative	37
AbrB	rapH	Negative	37
AbrB	rapK	Negative	37
AbrB	sacT	Negative	37
AbrB	ybaJ	Negative	37
AbrB	ycbJ	Negative	37
AbrB	yheJ	Negative	37
AbrB	ykuU	Negative	37
AbrB	yqeB	Negative	37
AbrB	yqfA	Negative	37
AbrB	yurI	Negative	37
AbrB	yvzA	Negative	37
AbrB	ywfB	Negative	37
AbrB	ybdG	Negative	60
AbrB	yczC	Negative	60
AbrB	yocD	Negative	61
AbrB	yqjZ	Negative	61
AbrB	yxiF	Negative	61
AbrB	yxxF	Negative	61
AbrB	appA	Negative	62

AbrB	appF	Negative	62
AbrB	yjcH	Negative	62
AbrB	yjiB	Negative	62
AbrB	yocS	Negative	62
AbrB	yueE	Negative	62
AbrB	ywaA	Negative	62
AbrB	acpA	Negative	104
AbrB	cspB	Negative	104
AbrB	tnrA	Negative	104
AbrB	veg	Negative	104
AbrB	yceC	Negative	104
AbrB	yceD	Negative	104
AbrB	yceK	Negative	104
AbrB	ykoF	Negative	104
AbrB	yktD	Negative	104
AbrB	ykuR	Negative	104
AbrB	ykzG	Negative	104
AbrB	ylmC	Negative	104
AbrB	ymfC	Negative	104
AbrB	yobQ	Negative	104
AbrB	yraM	Negative	104
AbrB	yceE	Negative	130
AbrB	yceF	Negative	130
AbrB	yokG	Negative	130
AbrB	yuiE	Negative	130
AbrB	yrdC	Negative	131
AbrB	hom	Negative	132
AbrB	mdr	Negative	132
AbrB	oppA	Negative	132
AbrB	ybgG	Negative	132
AbrB	ycgM	Negative	132
AbrB	ydeI	Negative	132
AbrB	yfhP	Negative	132
AbrB	ygaF	Negative	132
AbrB	yoaE	Negative	132
AbrB	yqgB	Negative	132
AbrB	yrpB	Negative	132
CcpA	glpQ	Negative	51
CcpA	glpT	Negative	51
CcpA	infC	Negative	51
CcpA	jag	Negative	51
CcpA	kbl	Negative	51

CcpA	rocB	Negative	51
CcpA	rplT	Negative	51
CcpA	rpmB	Negative	51
CcpA	rpmI	Negative	51
CcpA	spoIIIJ	Negative	51
CcpA	ybxF	Negative	51
CcpA	ycbF	Negative	51
CcpA	ycsA	Negative	51
CcpA	yesL	Negative	51
CcpA	yfiG	Negative	51
CcpA	yulD	Negative	51
CcpA	yvdF	Negative	51
CcpA	yvdH	Negative	51
CcpA	yvdJ	Negative	51
CcpA	msmR	Negative	52
CcpA	mtlA	Negative	52
CcpA	pmi	Negative	52
CcpA	sacA	Negative	52
CcpA	sacP	Negative	52
CcpA	ysbA	Negative	52
CcpA	epr	Negative	83
CcpA	lplA	Negative	83
CcpA	manP	Negative	83
CcpA	manR	Negative	83
CcpA	punA	Negative	83
CcpA	ydhM	Negative	83
CcpA	ydhT	Negative	83
CcpA	yncC	Negative	83
CcpA	yugN	Negative	83
CcpA	xsa	Negative	84
CcpA	ydeR	Negative	84
CcpA	ywcE	Negative	84
CcpA	ywcJ	Negative	84
CcpA	ywhE	Negative	84
CcpA	dhaS	Negative	86
CcpA	gutR	Negative	86
CcpA	ykoM	Negative	86
CcpA	yslB	Negative	86
CcpA	yurN	Negative	86
CcpA	citH	Negative	114
CcpA	galK	Negative	114
CcpA	pta	Negative	114

CcpA	yrzF	Negative	114
CcpA	yshB	Negative	114
CcpA	gcaD	Negative	115
CcpA	prs	Negative	115
CcpA	ybcF	Negative	115
CcpA	ylbN	Negative	115
CcpA	ytkP	Negative	115
CcpA	ytxD	Negative	115
CcpA	cccA	Negative	116
CcpA	cstA	Negative	116
CcpA	rocA	Negative	116
CcpA	rocC	Negative	116
CcpA	rocD	Negative	116
CcpA	rocE	Negative	116
CcpA	rocF	Negative	116
CcpA	yckB	Negative	116
CcpA	ydzA	Negative	116
CcpA	yodJ	Negative	116
CcpA	yojA	Negative	116
CcpA	ypmQ	Negative	116
CcpA	ytcP	Negative	116
CcpA	dctB	Negative	136
CcpA	sigI	Negative	136
CcpA	yydK	Negative	136
CodY	appC	Negative	55
CodY	appD	Negative	55
CodY	bioI	Negative	55
CodY	comGD	Negative	55
CodY	yjcG	Negative	55
CodY	ykvW	Negative	55
CodY	yqzE	Negative	55
CodY	yrkF	Negative	55
CodY	yuiA	Negative	55
CodY	yuxL	Negative	55
CodY	ywcE	Negative	55
CodY	yxiA	Negative	55
CodY	yxjG	Negative	55
ComK	pksI	Positive	22
ComK	ywfM	Positive	22
ComK	yydF	Positive	22
ComK	lexA	Positive	81
ComK	parC	Positive	81

ComK	parE	Positive	81
ComK	ruvA	Positive	81
ComK	ruvB	Positive	81
ComK	uvrA	Positive	81
ComK	uvrB	Positive	81
ComK	yhaM	Positive	81
ComK	yhaO	Positive	81
ComK	yqjY	Positive	81
ComK	yckE	Positive	138
ComK	yjbF	Positive	138
ComK	yqzG	Positive	138
ComK	bioA	Positive	139
ComK	bioB	Positive	139
ComK	bioD	Positive	139
ComK	bioF	Positive	139
ComK	bioI	Positive	139
ComK	bioW	Positive	139
ComK	clpP	Positive	139
ComK	dnaK	Positive	139
ComK	grpE	Positive	139
ComK	hrcA	Positive	139
ComK	metS	Positive	139
ComK	pabA	Positive	139
ComK	pabB	Positive	139
ComK	pabC	Positive	139
ComK	rph	Positive	139
ComK	slr	Positive	139
ComK	spoVB	Positive	139
ComK	yacF	Positive	139
ComK	ycgM	Positive	139
ComK	yisT	Positive	139
ComK	yokB	Positive	139
ComK	ytbQ	Positive	139
ComK	ywfB	Positive	139
ComK	ywrO	Positive	139
ComK	yxiF	Positive	139
ComK	pksC	Positive	141
ComK	rapK	Positive	141
ComK	ybdN	Positive	141
ComK	ybfO	Positive	141
ComK	ycbJ	Positive	141
ComK	yfmK	Positive	141

ComK	yvaW	Positive	141
ComK	yvaX	Positive	141
ComK	yvaY	Positive	141
ComK	yvaZ	Positive	141
ComK	yvbA	Positive	141
ComK	yvzA	Positive	141
ComK	yweA	Positive	141
ComK	ywqC	Positive	141
ComK	yybJ	Positive	141
ComK	yybK	Positive	141
ComK	yybL	Positive	141
ComK	yybM	Positive	141
ComK	yybN	Positive	141
ComK	yydG	Positive	141
ComK	yydH	Positive	141
ComK	yydI	Positive	141
ComK	yydJ	Positive	141
ComK	ycgL	Unknown	80
ComK	ydfL	Unknown	80
ComK	ydhH	Unknown	80
ComK	ywbO	Unknown	80
ComK	ywoB	Unknown	80
ComK	atpI	Unknown	82
ComK	hemL	Unknown	82
ComK	kinB	Unknown	82
ComK	rpsL	Unknown	82
ComK	rsbT	Unknown	82
ComK	sipV	Unknown	82
ComK	tnrA	Unknown	82
ComK	yabD	Unknown	82
ComK	yceK	Unknown	82
ComK	ydiQ	Unknown	82
ComK	ynfE	Unknown	82
ComK	yobQ	Unknown	82
ComK	yodF	Unknown	82
ComK	yogA	Unknown	82
ComK	yqeL	Unknown	82
ComK	yrvI	Unknown	82
ComK	yvzD	Unknown	82
ComK	ywqB	Unknown	82
ComK	yybB	Unknown	82
ComK	degS	Unknown	109

ComK	ilvB	Unknown	109
ComK	ilvC	Unknown	109
ComK	leuA	Unknown	109
ComK	leuB	Unknown	109
ComK	leuC	Unknown	109
ComK	leuD	Unknown	109
ComK	lonA	Unknown	109
ComK	nprE	Unknown	109
ComK	pycA	Unknown	109
ComK	sigH	Unknown	109
ComK	yfhB	Unknown	109
ComK	yjcH	Unknown	109
ComK	yjcI	Unknown	109
ComK	yjcJ	Unknown	109
ComK	yjcS	Unknown	109
ComK	yjdA	Unknown	109
ComK	yoaD	Unknown	109
ComK	yoaE	Unknown	109
ComK	yobK	Unknown	109
ComK	yokD	Unknown	109
ComK	ypdA	Unknown	109
ComK	ypfD	Unknown	109
ComK	yqgP	Unknown	109
ComK	yqjI	Unknown	109
ComK	yrpB	Unknown	109
ComK	yurK	Unknown	109
ComK	ywfH	Unknown	109
ComK	ywnF	Unknown	109
ComK	yxjG	Unknown	109
ComK	hemB	Unknown	140
ComK	ytaG	Unknown	140
Fur	fhuB	Negative	1
Fur	yetG	Negative	1
Fur	yxeC	Negative	8
Fur	yclM	Negative	46
Fur	ykuJ	Negative	46
Fur	fhuC	Negative	105
Fur	ykvW	Negative	122
Fur	ypbB	Negative	123
Fur	clpP	Unknown	58
Fur	dps	Unknown	58
Fur	gspA	Unknown	58

Fur	ydaN	Unknown	58
Fur	yflA	Unknown	58
Fur	ykzA	Unknown	58
Fur	yocB	Unknown	58
Fur	yqgZ	Unknown	58
Fur	yumB	Unknown	58
Fur	yvaA	Unknown	58
Fur	ywiE	Unknown	58
Fur	yxzF	Unknown	58
GerE	ylbD	Unknown	2
GerE	yisZ	Unknown	6
GerE	yitA	Unknown	6
GerE	yitB	Unknown	6
LexA	ymaC	Negative	9
LexA	yprA	Negative	9
LexA	yvsG	Negative	9
LexA	pcrB	Negative	128
PerR	dhbA	Unknown	8
PerR	dhbC	Unknown	8
PerR	feuA	Unknown	8
PerR	feuB	Unknown	8
PerR	feuC	Unknown	8
PerR	fhuB	Unknown	8
PerR	fhuD	Unknown	8
PerR	ybbA	Unknown	8
PerR	ydbN	Unknown	8
PerR	yfhA	Unknown	8
PerR	yfiY	Unknown	8
PerR	yfiZ	Unknown	8
PerR	yfmC	Unknown	8
PerR	yhfQ	Unknown	8
PerR	ykuN	Unknown	8
PerR	ykuO	Unknown	8
PerR	yoaJ	Unknown	8
PerR	yusV	Unknown	8
PerR	ywbL	Unknown	8
PerR	ywbM	Unknown	8
PerR	ywbN	Unknown	8
PerR	yxeB	Unknown	8
PerR	yxeC	Unknown	8
PerR	citZ	Unknown	29
PerR	tagB	Unknown	29

PerR	yurY	Unknown	29
PerR	dhbB	Unknown	137
PerR	dhbE	Unknown	137
PerR	dhbF	Unknown	137
PerR	ybbB	Unknown	137
PerR	yetG	Unknown	137
PerR	ykuP	Unknown	137
PerR	yuiI	Unknown	137
ResD	bglP	Positive	45
ResD	hemA	Positive	45
ResD	hemB	Positive	45
ResD	narG	Positive	45
ResD	narH	Positive	45
ResD	narI	Positive	45
ResD	narJ	Positive	45
ResD	ydbL	Positive	45
ResD	yhfO	Positive	45
ResD	yqxI	Positive	45
ResD	yueJ	Positive	45
ResD	yxiE	Positive	45
SigB	rsbR	Positive	7
SigE	yisZ	Positive	33
SigE	yitA	Positive	33
SigE	yitB	Positive	33
SigE	comER	Positive	72
SigE	sspB	Positive	72
SigE	ybaK	Positive	72
SigE	yhcN	Positive	72
SigE	yogA	Positive	72
SigE	yvlA	Positive	72
SigE	ylbI	Positive	74
SigE	sigG	Positive	75
SigE	xsa	Positive	75
SigE	cgeC	Positive	101
SigE	yfiA	Positive	101
SigE	yfnF	Positive	101
SigE	yfnG	Positive	101
SigE	yfnH	Positive	101
SigE	yqfQ	Positive	101
SigE	gapB	Positive	125
SigE	ykuD	Positive	125
SigE	oppB	Positive	127

SigF	csbX	Positive	50
SigF	yfhE	Positive	50
SigF	yfhF	Positive	50
SigH	ywfG	Positive	43
SigH	rapE	Positive	87
SigH	rapG	Positive	87
SigH	rapH	Positive	87
SigH	yqjZ	Positive	87
SigH	citG	Positive	120
SigW	appB	Positive	12
SigW	appC	Positive	12
SigW	appD	Positive	12
SigW	ywoF	Positive	12
SigW	dnaK	Positive	28
SigW	grpE	Positive	28
SigW	hrcA	Positive	28
SigW	ydgC	Positive	28
SigW	yjbI	Positive	28
SigW	ypuD	Positive	28
SigW	yrzG	Positive	28
SigW	yugM	Positive	28
SigW	yutM	Positive	28
SigW	ywdE	Positive	28
SigW	ywsA	Positive	28
SigW	yczC	Positive	56
SigW	ykuA	Positive	56
SigW	yugI	Positive	57
SigW	yuzE	Positive	98
SigW	yqgU	Positive	99
SigW	lytR	Positive	129
SigW	sunT	Positive	129
SigW	yncM	Positive	129
SigW	yxbC	Positive	129
SigW	yxbD	Positive	129
Spo0A	ywkC	Positive	31
Spo0A	yhjM	Unknown	32
Spo0A	yodL	Unknown	32
SpoIIID	dacB	Negative	10
SpoIIID	spoIIP	Negative	11
SpoIIID	yloB	Negative	17
SpoIIID	yoaN	Positive	69
SpoIIID	cotV	Unknown	13

SpoIIID	cotW	Unknown	13
TnrA	tuaA	Positive	66
TnrA	tuaB	Positive	66
TnrA	tuaC	Positive	66
TnrA	tuaD	Positive	66
TnrA	tuaF	Positive	66
TnrA	yrbD	Positive	66

Appendix 2

A list of all arrays and their assigned conditional categories. The column 'Array' refers to the experiment according to the source database (SMD, the first 3 arrays; GEO, the rest), while 'Category' refers to one of the 6 conditional categories, or to a control experiment.

Array	Category
C_26966	peroxide
C_26967	peroxide
C_26968	peroxide
GSM104507	control
GSM104511	control
GSM104514	control
GSM104517	DNA
GSM104518	DNA
GSM104519	DNA
GSM104520	DNA
GSM104521	DNA
GSM104522	DNA
GSM104523	DNA
GSM104524	DNA
GSM104525	DNA
GSM104526	control
GSM104527	control
GSM104528	control
GSM104727	DNA
GSM104728	DNA
GSM104729	DNA
GSM104730	DNA
GSM104731	DNA
GSM104732	DNA
GSM104733	DNA
GSM104734	DNA
GSM104735	DNA
GSM105249	sporulation
GSM105250	sporulation
GSM105251	sporulation
GSM105252	DNA

GSM105253	DNA
GSM105254	DNA
GSM105255	DNA
GSM105256	DNA
GSM105257	DNA
GSM105258	DNA
GSM105259	DNA
GSM105260	DNA
GSM105261	control
GSM105262	control
GSM105263	control
GSM105264	DNA
GSM105265	DNA
GSM105266	DNA
GSM105267	DNA
GSM105268	DNA
GSM105269	DNA
GSM105270	DNA
GSM105271	quorum response
GSM105272	DNA
GSM105273	DNA
GSM105280	quorum response
GSM105281	quorum response
GSM105282	quorum response
GSM105283	quorum response
GSM105284	quorum response
GSM105285	quorum response
GSM105286	quorum response
GSM105287	quorum response
GSM105288	quorum response
GSM105289	quorum response
GSM105290	quorum response
GSM105326	control
GSM105327	control
GSM105328	control
GSM105329	quorum response
GSM105330	DNA
GSM105331	DNA
GSM105332	DNA
GSM105345	DNA
GSM105346	quorum response
GSM105347	DNA

GSM105348	DNA
GSM105349	quorum response
GSM105350	DNA
GSM105351	DNA
GSM105352	quorum response
GSM105353	DNA
GSM105354	quorum response
GSM105355	quorum response
GSM105356	quorum response
GSM105357	quorum response
GSM105358	quorum response
GSM105359	control
GSM105360	control
GSM105361	control
GSM105416	DNA
GSM105417	DNA
GSM105418	DNA
GSM105419	control
GSM105420	control
GSM105421	control
GSM105422	DNA
GSM105423	DNA
GSM105424	DNA
GSM105425	DNA
GSM105426	DNA
GSM105427	DNA
GSM27217	sporulation
GSM27218	sporulation
GSM27219	sporulation
GSM27278	sporulation
GSM27279	sporulation
GSM27280	sporulation
GSM27281	sporulation
GSM27282	sporulation
GSM27283	sporulation
GSM27285	sporulation
GSM27286	sporulation
GSM27288	sporulation
GSM27514	sporulation
GSM27515	sporulation
GSM27516	sporulation
GSM49830	control

GSM49975	control
GSM49998	phosphate
GSM50000	phosphate
GSM50002	phosphate
GSM50003	phosphate
GSM50008	control
GSM50378	phosphate
GSM50472	phosphate
GSM51432	phosphate
GSM51433	phosphate
GSM51434	phosphate
GSM51435	control
GSM51436	phosphate
GSM51437	phosphate
GSM51438	phosphate
GSM51439	phosphate
GSM51440	phosphate
GSM51441	control
GSM51442	phosphate
GSM51443	phosphate
GSM51444	phosphate
GSM51445	phosphate
GSM51446	control
GSM51447	control
GSM51448	phosphate
GSM51449	phosphate
GSM51450	phosphate
GSM51451	control
GSM51452	control
GSM51453	phosphate
GSM51454	phosphate
GSM51455	phosphate
GSM51456	phosphate
GSM51457	control
GSM51458	control
GSM51459	phosphate
GSM51460	phosphate
GSM51461	phosphate
GSM70016	peroxide
GSM70017	peroxide
GSM70018	peroxide
GSM70019	peroxide

GSM70020	peroxide
GSM70021	peroxide
GSM70022	peroxide
GSM70023	peroxide
GSM70024	heat
GSM70025	heat
GSM70026	heat
GSM70027	heat
GSM70028	peroxide
GSM70029	heat
GSM70030	peroxide
GSM70031	heat
GSM70032	heat
GSM70033	peroxide
GSM70034	heat
GSM70035	peroxide
GSM70036	heat
GSM70037	heat
GSM70038	peroxide
GSM70039	peroxide
GSM70040	peroxide
GSM70041	peroxide
GSM70042	peroxide
GSM70043	peroxide
GSM70044	peroxide
GSM70045	peroxide
GSM70046	peroxide
GSM70047	peroxide
GSM70048	peroxide
GSM70049	peroxide
GSM70050	peroxide
GSM70051	peroxide
GSM70052	heat
GSM70053	peroxide
GSM70054	heat
GSM70055	peroxide
GSM70056	heat
GSM70057	peroxide
GSM70058	peroxide
GSM70059	heat
GSM70060	peroxide
GSM70061	peroxide

GSM70062	peroxide
GSM70063	peroxide
GSM70064	peroxide
GSM70065	peroxide
GSM70066	peroxide
GSM70067	peroxide
GSM70068	peroxide
GSM70069	peroxide
GSM70070	peroxide
GSM70071	peroxide
GSM70072	peroxide
GSM70073	peroxide
GSM70074	peroxide
GSM70075	peroxide
GSM70076	peroxide
GSM70077	peroxide
GSM70078	peroxide
GSM70079	peroxide
GSM70080	peroxide
GSM70081	peroxide
GSM70082	heat
GSM70083	peroxide
GSM70084	heat
GSM70085	peroxide
GSM70086	peroxide
GSM70087	peroxide
GSM70088	peroxide
GSM70089	peroxide
GSM70090	peroxide
GSM70091	peroxide
GSM70092	peroxide
GSM70093	peroxide

Appendix 3

A list of predicted intra-operon motifs

Module	Gene	Operon	Hertzberg p-value	Predicted Regulators	Upstream operon gene belongs to a different module	Predicted by Hoon *et al.*
module 63	acoB	acoABCL	0.999303	motif_14	1	0
module 64	acoB	acoABCL	0.999303	motif_14	1	0
module 55	appC	appDFABC	0.993966	CodY	1	0
module 85	araP	araABDLMNPQ-abfA	0.974341	CcpA	0	1
module 55	bioI	bioWAFDBI-ytbQ	0.992156	CodY	1	0
module 65	cgeB	cgeAB	0.977814	motif_75	1	0
module 109	comEB	comEABC	0.957916	ComK	1	0
module 110	comEB	comEABC	0.957916	ComK	1	0
module 139	comEB	comEABC	0.957916	ComK	1	0
module 22	comEB	comEABC	0.957916	ComK	1	0
module 34	degU	degSU	0.999727	motif_187	1	1
module 55	dppC	dppABCDE	0.984839	CodY	1	0
module 67	flhO	usd-spoIIID-mbl-flhOP	0.975413	motif_106	1	1
module 119	folD	nusB-folD	0.998672	PurR	1	0
module 113	glpK	glpFK	0.999535	CcpA	1	0
module 115	glpK	glpFK	0.999535	CcpA	1	0
module 135	glpK	glpFK	0.999535	CcpA	1	0
module 136	glpK	glpFK	0.999535	CcpA	1	0
module 52	glpK	glpFK	0.999535	CcpA	1	0
module 67	glpK	glpFK	0.957679	motif_106	1	0
module 86	glpK	glpFK	0.999535	CcpA	1	0
module 29	hemC	hemAXCDBL	0.968333	PerR	1	0

module 34	hemL	hemAXCDBL	0.956952	motif_187	1	0
module 82	hemL	hemAXCDBL	0.999537	ComK	1	0
module 114	iolD	mmsA-iolBCDEF-idh-iolHI-fbaB	0.972834	CcpA	1	1
module 64	mbl	usd-spoIIID-mbl-flhOP	0.970264	motif_14	0	1
module 102	mmgD	mmgABCDE-yqiQ	0.972412	SigE	0	0
module 71	mmgD	mmgABCDE-yqiQ	0.972412	SigE	0	0
module 115	nupC	deoR-dra-nupC-pdp	0.994555	CcpA	1	0
module 64	nupC	deoR-dra-nupC-pdp	0.986046	motif_14	1	0
module 86	nupC	deoR-dra-nupC-pdp	0.994555	CcpA	1	0
module 127	oppB	oppABCDF	0.96366	SigE	1	0
module 83	punA	drm-punA	0.96047	CcpA	0	0
module 119	purM	purEKBCSQLFMNHD	0.988785	PurR	1	0
module 131	qcrC	qcrABC	0.95309	AbrB	1	0
module 134	rbsC	rbsRKDACB	0.973193	CcpA	1	0
module 136	rbsC	rbsRKDACB	0.973193	CcpA	1	0
module 86	rbsC	rbsRKDACB	0.973193	CcpA	1	0
module 51	rocB	rocABC	0.974141	CcpA	1	0
module 115	rplT	infC-rpmI-rplT-ysdA	0.960696	CcpA	1	0
module 40	rpsF	yyaF-rpsF-ssb-rpsR	0.999422	motif_2	0	1
module 40	rpsR	yyaF-rpsF-ssb-rpsR	0.953967	motif_2	0	0
module 82	rsbT	rsbRSTUVW-sigB-rsbX	0.951539	ComK	0	0
module 82	sipV	yhjE-sipV	0.998623	ComK	0	0
module 29	tagB	tagAB	0.99933	PerR	1	0
module 34	trpF	trpEDCFBA-hisC-tyrA-aroE	0.963127	motif_187	0	0
module 139	yacF	pabBAC-sul-folBK-yazB-yacF-lysS	0.982793	ComK	0	0
module 91	yacK	ctsR-mcsAB-clpC-radA-yacK	0.967746	motif_53	0	0
module 132	yceD	yceCDEFGH	0.999217	AbrB	1	0

module 63	yfhE	yfhFED	0.992692	motif_14	1	1
module 101	yfiA	malA-yfiA-malP	0.968736	SigE	0	0
module 34	yfnD	yfnHGFED	0.964504	motif_187	1	0
module 81	yhaM	yhaONM	0.96086	ComK	0	1
module 67	ylxF	fla-che	0.952389	motif_106	1	0
module 40	ylxR	ylxS-nusA-ylxRQ-infB-ylxP-rbfA	0.956298	motif_2	0	0
module 91	yoxB	yoxCB-yoaA	0.970319	motif_53	1	0
module 92	yqcK	arsR-yqcK-arsBC	0.999209	motif_53	0	1
module 37	yqfA	yqeZ-yqfA-yqfB	0.959228	AbrB	1	0
module 60	yqfA	yqeZ-yqfA-yqfB	0.959228	AbrB	1	0
module 55	yqzE	comGABCDEFG-yqzE	0.99129	CodY	1	1
module 82	yrvI	relA-yrvI	0.952573	ComK	0	1
module 121	ytxH	ytxGHJ	0.960953	motif_260	1	0
module 38	yurG	yurHG	0.961286	motif_3	0	1
module 63	yvfM	yvfKLM-lacA-yvfO	0.997027	motif_14	0	0
module 92	ywcH	nfrA-ywcH	0.958723	motif_53	0	1
module 45	yxiE	bglPH-yxiE	0.990092	ResD	1	0