Inferring the transcriptional network of *Bacillus subtilis*

Authors:

Abeer Fadda[1], Ana Carolina Fierro[1], Karen Lemmens[1], Pieter Monsieurs[2], Kristof Engelen[1], Kathleen Marchal[1]*

[1]Department of Microbial and Molecular Systems, KULeuven, Kasteelpark Arenberg 20, 3001 Heverlee, Belgium

[2]Institute for Environment, Health and Safety, Belgian Nuclear Research Center, SCK•CEN, Boeretang 200, B-2400 Mol, Belgium

* Corresponding author

Abstract

The adaptation of bacteria to the vigorous environmental changes they undergo is crucial to their survival. They achieve this adaptation partly via the intricate regulation of the transcription of their genes. In this study, we infer the transcriptional network of the Gram-positive model organism *Bacillus subtilis.* We use a data integration workflow, exploiting both motif and expression data, towards the generation of condition-dependent transcriptional modules. In building the motif data, we rely on both known and predicted information. Known motifs were derived from DBTBS while predicted motifs were generated by a *de novo* motif detection method that utilizes comparative genomics. 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.

Introduction

Understanding the global properties of transcriptional networks in bacteria is key to an improved understanding of the versatile and adaptive lifestyles of these organisms. So far, most studies have been performed on the model organism *Escherichia 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[1,2,3,4,5,6]. Few studies, however, addressed the "dynamic" network, either with respect to time[7] or with respect to environmental conditions[8,9,10]. Indeed, most genes undergo conditional regulation rather than a constitutive one.

From this dynamic viewpoint, 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[11,12]. 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[9]. Integrating different data types not only results in predicting new interactions with more confidence than when relying on a single data type, but it also allows for a more global view on the network. The information on the network structure was derived from motif information, 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.

Results

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 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[13]), 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 a *de novo* motif detection method to predict motifs for regulators with yet unknown binding sites. In this case, the direct relation between a regulator and these binding sites is unknown and needs to be inferred as well. The *de novo* motif detection strategy we used is based on a combination of phylogenetic footprinting with the concept of co-regulation; however, co-regulation in this case relies on sequence data alone so as to generate information that is completely independent from the expression compendium. The basic idea behind this approach is that motifs that occur within

evolutionary conserved regions in the promoters of different species (phylogenetic footprinting) and that are shared by at least two genes of the same species (co-regulation) are most likely to be functional [14,15,16]. Potential regulatory motifs were thus identified as those sequence regions conserved over four different *Bacillus* species and shared by at least two different *B. subtilis* genes. According to this strategy 159 candidate motifs were selected. Note that by using this double criterion of phylogenetic conservation and co-regulation, we could reduce the motif search space in a sensible fashion (the total number of predicted motif sites dropped from 7543 at the phylogenetic step, to 393 after the final motif extraction step). On the other hand, motifs of regulators that have only one target gene in *B. subtilis* are missed. Also, motif instances in genes that do not have a conserved regulation, and instances present in genes that are not conserved across all four *Bacillus* species, are missed. The latter instances could be recovered by performing a final genome wide screening using the retrieved motif models. Details can be found in the materials and methods section.
For network reconstruction, we used the data integration framework 'DISTILLER'. DISTILLER uses a strategy based on itemset mining to integrate expression with motif data[10]. 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.

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. A detailed description of the modules can be found at the supporting webpage http://homes.esat.kuleuven.be/~afadda/supporting/modules.html. 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 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 1). A detailed list of the interactions predicted per regulator can be found in the supporting table S3. The highest number of predicted novel targets was for the regulators with global functions, such as competence (ComK) and catabolite repression (CcpA, AbrB) (Table 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 1. 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

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 1). 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 table S3.

Prediction of regulons for yet uncharacterized regulators

For *B. subtilis* there still exists a pool of regulators with largely unknown regulons. They comprise either proteins with experimentally verified regulatory functions and unknown motifs, or entirely uncharacterized hypothetical regulators. Regulons for this pool of regulators were derived from the modules of the *de novo* motifs. Integrating the predicted *de novo* motifs (see the results section on network construction) with expression data, allowed us to assign 13 of our *de novo* motifs to at least

one module (Table 2). 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 putative regulators, we used several criteria. First, we compared the motif models with those present in DBTBS for known regulators (45 models). Second, we tested whether the genes containing those motifs are enriched for targets of a specific regulator. The regulators that fulfilled both criteria were considered as potential matches to our *de novo* motifs. Of the 13 motifs that had a module in our study, motif_12 matched with Fur, and motif_260 matched with SinR. Ideally, if the match based on sequence and functional data between the novel motif and a known motif was true, we expect the known regulator (in this case Fur or SinR) to be assigned to the same modules of the novel motif (motif_12 or motif_260). This was the case for the Fur motif (8 out of 11 modules assigned to Fur are also assigned to motif_12), but not for SinR (two modules were assigned to motif_260, and none to SinR). The discrepancy can be explained by the variation in the motif models for SinR as detected by our approach and as present in DBTBS, resulting in differences in the scores of the motif screening.

For all other motifs not assigned to a known regulator, 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[17], 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 and supplementary table S1 and figure S2). Secondly, bacterial regulators are often auto-regulated: 55% of the regulators in *E. coli*[3] and 42% in *B. subtilis* (this study) are known to be auto-regulated. 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 2) 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[18]. 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[19]. 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[20].

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. Microarray experiments under each category assess either mutants and/or conditions related to the assigned category (see supporting table S5 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 3).
For example, regulators whose modules are enriched for 'DNA' include LexA, the best known regulator of DNA damage-inducible genes[21], and SigD, a regulator of the autolytic response among other functions[22]. 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[23]; 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)[24]. PerR also exhibits 'activity' in the peroxide category, as would be expected for the repressor of the peroxide regulon[25].
The "phosphate" category includes for the most part arrays related to phosphate starvation. Starvation is a known trigger for sporulation [26], 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[27] with an activation role in the global regulation of aerobic and anaerobic respiration[28], is active in this category too.
For the condition 'sporulation', SpoIIID, GerE, SigK, and SigE, all known regulators of sporulation[29] 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[30] and the early-stage sporulation regulators, Spo0A and SigH[29].
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[31]. 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[32] and their encoding genes undergo a decreased expression[33]. Our results suggest that a similar process to the one in *E. coli* could be taking place in *B. subtilis*.

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;[34]. 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 supporting table S4). Figure 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 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 4A). The transcription profile of these two genes is indeed very different from that of the rest of the operon genes (Figure 4B).

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[35] 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[25]. 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[36], it is not surprising to find them as regulators of a common complex regulon.

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[1,37], others include additional criteria such as the vast number of environmental conditions to which these regulators respond[3,38] 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 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 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[3]. 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 3. Functions of the global regulators

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

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 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[39], a signaling mechanism known to influence the development of competence in a cell[40]. 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[29]. 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; oftentimes 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[41,42]. 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.

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[43] (De Hoon *et al.* 2004) and the sporulation initiation genetic network[44]. 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-dependent, 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[45]. 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 supporting table S4). 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[46,47]. 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*.

Materials and methods

De novo motif detection

For the detection of novel motifs we relied on a *de novo* motif detection strategy, using the genome wide application of the methodology described in Monsieurs *et al.*[14].

The following four *Bacillus* species were used for the first phylogenetic footprinting step (supporting figure S1) as they exhibited an optimal phylogenetic relatedness i.e. not too closely related to be uninformative and not too distantly related to have altered their mechanism of transcriptional regulation[48]): *B. subtilis* (AL009126), *B. anthracis* (NC_007530), *B. licheniformis* (NC_006322) and *B. cereus* (NC_004722). To increase the reliability of the footprint results, we only searched for motifs that were conserved in the promoter regions of genes present in all four species; by pair-wise reciprocal BLAST best hits we identified 1943 gene sets that included all four *Bacillus* species. Intergenic sequences upstream of the gene sets were extracted. Intergenic sequences are the non-coding sequences upstream of the annotated ATG in GenBank. Gene sets for which the intergenic sequences were at least 40 bp were withheld, as they are the most likely to contain regulatory regions in their promoter region. This resulted in 1284 intergenic gene sets.

For each gene set, intergenic sequences were locally aligned (step 1 in supporting figure S1) using the stochastic algorithm BlockSampler[14] with the following settings: we set *B. subtilis* as the reference sequence and used species specific third-order background models based on all intergenic sequences. Being a stochastic algorithm, BlockSampler was run 100 times per orthologous intergenic set using default parameters (searching on the plus strand, prior = 0.2, consensus score threshold = 1.3, minimum block width = 8), resulting in a list of 100 conserved regions (blocks). Redundant blocks were filtered from the list as follows: blocks overlapping for more than 75% were identified and listed. For each set of redundant blocks a representative was identified as the one with the best log likelihood score. All representatives were subsequently listed and ranked according to their log likelihood score: the 6 top scoring representatives were retained for each orthologous intergenic set.

In a subsequent step, the blocks selected from all intergenic sets were mutually compared to identify conserved regions (blocks) that are common between different orthologous intergenic sets (step 2 in supporting figure S1). To this end, each block was converted into a position specific weight matrix and the resulting PWMs were mutually compared using the algorithm BlockAligner (based on the Kullback-Leibler distance;[14]. The algorithm was set to report PWMs (or blocks) that significantly align over a minimum of 6 bp. Significance was assessed by using 100 randomizations per alignment. Pair-wise alignments with a p-value < 0.0001 were considered significant. The scores of the pair-wise alignments (i.e. p-values) were used as input for a graph-based fuzzy clustering algorithm[49]. Ten clustering solutions were obtained and merged, and a probability cut off value of 0.5 was used to obtain tight clusters. This resulted in 300 clusters, containing 2–12 blocks per cluster. A cluster, thus, consists of different blocks, originating from different orthologous intergenic sets that exhibit mutual similarity. The regions of such blocks that are conserved over different orthologous intergenic sets (or conserved cores) thus correspond to putative regulatory motifs. To extract these cores we selected the *B. subtilis* representative sequence for each block in a cluster and aligned them using a local multiple alignment strategy[50]. This resulted in a total of 159 reliable, putative novel motifs and their corresponding motif models. For the evaluation of the motif reduction step by comparative genomics we refer to the supplementary information.

Constructing a motif compendium

Both the *de novo* motif models 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.*[51]. 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 translation start site.

Construction of the microarray compendium

A compendium of 231 publicly available microarrays were downloaded from the Stanford Microarray Database[52] and Gene Expression Omnibus[53], derived from 5 independent studies and performed on 10 different platforms. An overview of the compendium can be found in Table 4. 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 4. 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

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 $\leq$ 0.001, and instances documented in DBTBS (including intra-operon regulation) were set to one. For DISTILLER we used the following parameter settings based on a parameter sweep: 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. The details of the parameter sweep are shown in the supporting material in section “Selecting parameter values for DISTILLER” and table S1 and figure S2. Post-processing and seed-extension were performed as described in Lemmens *et al.*[10]

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 recall 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).

Associating *de novo* motifs with potential regulators

To associate our *de novo* motifs with known regulators, we first calculated the enrichment of the gene sets containing the novel motifs for targets of a specific regulator, using the hypergeometric distribution. This was performed for all the 120 regulators listed in DBTBS, and all tests with a p-value ≤ 0.05 were retained. Secondly, we compared the motif models of the *de novo* motifs to all models present in DBTBS (45 motif models), using the Kullback-Leibler distance measure with the software TOMTOM[54]. Comparisons were done for a minimum overlap of 5 bp, and all tests with an E-value < 1 were retained. Assignment of a regulator to a *de novo* motif was considered likely when both criteria mentioned above were fulfilled. This resulted in the assignment of 8 *de novo* motifs to a known regulator.

For the rest of the motifs, we assigned potential regulators as follows:

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[55]. 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). A distance of 2500 bases gave the best combination of recall (0.69) and precision (0.25) when tested against known TF-targets (see supporting table S2 and figure S3). This distance was used as an estimate for the minimal distance within which a regulator and its corresponding closest target gene are co-localized.

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[56]. A p-value ≤ 0.05 was considered significant. The same method was used to calculate the condition enrichment using the six categories described before.

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 supporting table S4.

Acknowledgements

The authors wish to thank Anagha Joshi for her help with the motif detection. This work was supported by 1) Belgian Technical Cooperation; 2) KUL: (GOA/08/011), CoE EF/05/007 SymBioSys; 3)

CREA/08/023; 4) IWT: SBO-BioFrame; 5) IUAP P6/25 (BioMaGNet); 6) FWO: G.031805, G.0329.09; 7) HSF RGY0079/207C

References

1 S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, *Nat. Genet.,* 2002, **31,** 1, pp. 64-68.
2 H. W. Ma, J. Buer, and A. P. Zeng, *BMC. Bioinformatics,* 2004, **5,** pp. 199.
3 A. Martínez-Antonio and J. Collado-Vides, *Curr. Opin. Microbiol.,* 2003, **6,** 5, pp. 482-9.
4 S. Balaji, M. M. Babu, and L. Aravind, *J. Mol. Biol.,* 2007, **372,** 4, pp. 1108-22.
5 J. Ernst, Q. K. Beg, K. A. Kay, G. Balázsi, Z. N. Oltvai, and Z. Bar-Joseph, *PLoS. Comput. Biol.,* 2008, **4,** 3, pp. e1000044.
6 H. Yu and M. Gerstein, *Proc. Natl. Acad. Sci. U. S. A.,* 2006, **103,** 40, pp. 14724-31.
7 J. Ernst, O. Vainas, C. T. Harbison, I. Simon, and Z. Bar-Joseph, *Mol. Syst. Biol.,* 2007, **3,** pp. 74.
8 N. M. Luscombe, M. M. Babu, H. Yu, M. Snyder, S. A. Teichmann, and M. Gerstein, *Nature,* 2004, **431,** 7006, pp. 308-12.
9 K. Lemmens, T. Dhollander, T. De Bie, P. Monsieurs, B. De Moor, J. Collado-Vides, K. Engelen, and K. Marchal, *Ann. N. Y. Acad. Sci.,* 2008,
10 K. Lemmens, T. De Bie, T. Dhollander, S. C. De Keersmaecker, I. M. Thijs, G. Schoofs, A. De Weerdt, B. De Moor, J. Vanderleyden, J. Collado-Vides, K. Engelen, and K. Marchal, *Genome. Biol.,* 2009, **10,** 3, pp. R27.
11 J. Errington, *Rev. Microbiol.,* 2003, **1,** pp. 117-126.
12 D. Lopez, H. Vlamakis, and R. Kolter, *FEMS. Microbiol. Rev.,* 2009, **33,** 1, pp. 152-63.
13 Y. Makita, M. Nakao, N. Ogasawara, and K. Nakai, *Nucleic. Acids. Res.,* 2004, **32,** Database issue, pp. D75-7.
14 P. Monsieurs, G. Thijs, A. A. Fadda, S. C. De Keersmaecker, J. Vanderleyden, B. De Moor, and K. Marchal, *BMC Bioinformatics,* 2006, **7,** pp. 160.
15 G. D. Stormo and G. W. Hartzell, *Proc. Natl. Acad. Sci. U. S. A.,* 1989, **86,** 4, pp. 1183-7.
16 T. Wang and G. D. Stormo, *Bioinformatics,* 2003, **19,** 18, pp. 2369-80.
17 K. Tan, L. A. McCue, and G. D. Stormo, *Genome. Res.,* 2005, **15,** 2, pp. 312-20.
18 B. R. Belitsky, M. C. Gustafsson, A. L. Sonenshein, and C. Von Wachenfeldt, *J. Bacteriol.,* 1997, **179,** 17, pp. 5448-57.
19 A. Burkovski, *Appl. Microbiol. Biotechnol.,* 2002, **58,** 3, pp. 265-274.
20 G. E. McDonnell and D. J. McConnell, *J. Bacteriol.,* 1994, **176,** 18, pp. 5831-5834.
21 M. F. Wojciechowski, K. R. Peterson, and P. E. Love, *J. Bacteriol.,* 1991, **173,** 20, pp. 6489-98.
22 L. M. Márquez, J. D. Helmann, E. Ferrari, H. M. Parker, G. W. Ordal, and M. J. Chamberlin, *J. Bacteriol.,* 1990, **172,** 6, pp. 3435-43.
23 K. W. Bayles, *Trends. Microbiol.,* 2003, **11,** 7, pp. 306-11.
24 M. S. Turner and J. D. Helmann, *J. Bacteriol.,* 2000, **182,** 18, pp. 5202-10.
25 N. Bsat, A. Herbig, L. Casillas-Martinez, P. Setlow, and J. D. Helmann, *Mol. Microbiol.,* 1998, **29,** 1, pp. 189-98.
26 A. L. Sonenshein, *Curr. Opin. Microbiol.,* 2000, **3,** 6, pp. 561-566.
27 S. M. Birkey, W. Liu, X. Zhang, M. F. Duggan, and F. M. Hulett, *Mol. Microbiol.,* 1998, **30,** 5, pp. 943-53.
28 M. M. Nakano, Y. Zhu, M. Lacelle, X. Zhang, and F. M. Hulett, *Mol. Microbiol.,* 2000, **37,** 5, pp. 1198-207.
29 Piggot, PJ and Hilbert, DW, *Curr. Opin. Microbiol.,* 2004, **7,** 6, pp. 579-584.
30 P. Eichenberger, M. Fujita, S. T. Jensen, E. M. Conlon, D. Z. Rudner, S. T. Wang, C. Ferguson, K. Haga, T. Sato, J. S. Liu, and R. Losick, *PLoS. Biol.,* 2004, **2,** 10, pp. e328.
31 J. Y. Cha, J. S. Lee, J. I. Oh, J. W. Choi, and H. S. Baik, *Biochem. Biophys. Res. Commun.,* 2008, **366,** 2, pp. 281-7.
32 B. D. Davis, S. M. Luger, and P. C. Tai, *J. Bacteriol.,* 1986, **166,** 2, pp. 439-45.
33 J. H. Baek and S. Y. Lee, *J. Microbiol. Biotechnol.,* 2007, **17,** 2, pp. 244-52.
34 E. Laing, V. Mersinias, C. P. Smith, and S. J. Hubbard, *Genome. Biol.,* 2006, **7,** 6, pp. R46.

35 R. M. Gutiérrez-Ríos, D. A. Rosenblueth, J. A. Loza, A. M. Huerta, J. D. Glasner, F. R. Blattner, and J. Collado-Vides, *Genome. Res.,* 2003, **13,** 11, pp. 2435-43.
36 D. D. Agranoff and S. Krishna, *Mol. Microbiol.,* 1998, **28,** 3, pp. 403-12.
37 M. M. Babu and S. A. Teichmann, *Nucleic. Acids. Res.,* 2003, **31,** 4, pp. 1234.
38 S. Gottesman, *Annual. Reviews. in. Genetics.,* 1984, **18,** 1, pp. 415-441.
39 L. W. Hamoen, G. Venema, and O. P. Kuipers, *Microbiology,* 2003, **149,** Pt 1, pp. 9-17.
40 P. Tortosa, L. Logsdon, B. Kraigher, Y. Itoh, I. Mandic-Mulec, and D. Dubnau, *J. Bacteriol.,* 2001, **183,** 2, pp. 451-60.
41 N. Schwartz, R. Cohen, D. Ben-Avraham, A. L. Barabási, and S. Havlin, *Phys. Rev. E. Stat. Nonlin. Soft. Matter. Phys.,* 2002, **66,** 1 Pt 2, pp. 015104.
42 H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, A. L. Barabasi, and O. others, *Nature,* 2000, **407,** 6804, pp. 651-653.
43 M. J. de Hoon, Y. Makita, S. Imoto, K. Kobayashi, N. Ogasawara, K. Nakai, and S. Miyano, *Bioinformatics,* 2004, **20 Suppl 1,** pp. i101-8.
44 H. de Jong, J. Geiselmann, C. Hernandez, and M. Page, *Bioinformatics,* 2003, **19,** 3, pp. 336-44.
45 M. J. De Hoon, S. Imoto, K. Kobayashi, N. Ogasawara, and S. Miyano, *Pac. Symp. Biocomput.,* 2004, pp. 276-87.
46 K. J. Boor, *PLoS. Biol.,* 2006, **4,** 1, pp. e23.
47 M. J. Kazmierczak, M. Wiedmann, and K. J. Boor, *Microbiol. Mol. Biol. Rev.,* 2005, **69,** 4, pp. 527-543.
48 K. Marchal, S. De Keersmaecker, P. Monsieurs, N. van Boxel, K. Lemmens, G. Thijs, J. Vanderleyden, and B. De Moor, *Genome. Biol.,* 2004, **5,** 2, pp. R9.
49 A. Joshi, Y. Van de Peer, and T. Michoel, *Bioinformatics,* 2008, **24,** 2, pp. 176.
50 G. Thijs, K. Marchal, M. Lescot, S. Rombauts, B. De Moor, P. Rouzé, and Y. Moreau, *J. Comput. Biol.,* 2002, **9,** 2, pp. 447-64.
51 L. Hertzberg, O. Zuk, G. Getz, and E. Domany, *J. Comput. Biol.,* 2005, **12,** 3, pp. 314-330.
52 J. Demeter, C. Beauheim, J. Gollub, T. Hernandez-Boussard, H. Jin, D. Maier, J. C. Matese, M. Nitzberg, F. Wymore, Z. K. Zachariah, and O. others, *Nucleic. Acids. Res.,* 2007, **35,** Database issue, pp. D766.
53 R. Edgar, M. Domrachev, and A. E. Lash, *Nucleic. Acids. Res.,* 2002, **30,** 1, pp. 207.
54 S. Gupta, J. A. Stamatoyannopoulos, T. L. Bailey, and W. S. Noble, *Genome. Biol.,* 2007, **8,** 2, pp. R24.
55 S. K. Kummerfeld and S. A. Teichmann, *Nucleic. Acids. Res.,* 2006, **34,** Database issue, pp. D74-81.
56 R.A. Fisher, *Statistical Methods for Research Workers,* Macmillan Pub Co., 15 edn.,1970.

Figure legends

Figure 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.

Figure 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.

Figure 3. Conditional dependency of the network. Rows: regulators or predicted motifs assigned to modules. Columns: condition categories (see supporting table S4 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.

Figure 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).

Figure 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.

Figure 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.

Table 2. Regulons predicted by our integrative analysis. 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 (TFs with an *), except in two cases (**) where the regulator was assigned based on other criteria explained in the text.