Comparative systems biology of the sporulation initiation network in prokaryotes

Michiel de Hoon and Dennis Vitkup

Columbia University, Center for Computational Biology and Bioinformatics, New York NY 10032, United States
mdehoon@c2b2.columbia.edu

Abstract. Many years of experimental and computational molecular biology of model organisms such as *Escherichia coli* and *Saccharomyces cerevisiae* has elucidated the gene regulatory network in these organisms. Relatively little is known about gene regulation in species other than the model organisms, whether gene regulatory networks are conserved, and to what degree our knowledge based on model organisms reflects biological networks occurring in nature as a whole.

In this paper, we describe a first attempt to understand the gene regulatory network in lesser-known organisms, using our knowledge of gene regulation in a well-understood model organism. Such an extrapolation is particularly valuable in the study of disease-causing infectious agents, as well as other organisms that are difficult to grow or handle in a laboratory environment. In addition, comparative systems biology can identify which parts of biological networks are poorly understood and are therefore promising venues for further experimental research.

We analyze the gene regulatory network responsible for the initiation of sporulation in fourteen target organisms, using *Bacillus subtilis* as the model organism. Instead of focusing on individual transcription factor binding sites, we devise a scoring function that takes into account the effect of multiple transcription factors binding to the regulatory region. Whereas the core gene regulatory network appears to be conserved, the degree of conservation decreases rapidly for more remote organisms, as well as for regulatory relations in the periphery of the network. Our work shows that gene regulation is still poorly understood in species other than the model organisms.

1 Introduction

In the post-genomic era, one of the major goals of molecular biology is to understand the gene regulatory network that drives the expression of genes depending on cellular conditions. Gene regulation is mediated by transcription factors, proteins that stimulate or repress the expression of genes by binding to the regulatory DNA sequence in their upstream region. Transcription factors recognize specific motifs in the DNA sequence, enabling them to differentially regulate genes based on their respective regulatory code. In many cases, the DNA motifs recognized by a specific transcription factor are similar to each other, which allows us to define a consensus binding motif that can be used to detect potential transcription factor binding sites in a DNA sequence.

For well-studied organisms such as *Escherichia coli* and *Saccharomyces cerevisiae*, a large number of transcription factors, their regulated genes, and the DNA binding sites have been found experimentally by DNA footprinting, disrupting transcription factors, mutating DNA binding sites, and primer extension experiments. Computationally, we can find transcription factor binding sites using comparative genomics, in which the upstream DNA sequences of homologous genes in nearby species are aligned to find conserved motifs. This approach relies on the assumption that the regulatory network is conserved between nearby species, which may or may not be true.

Whereas the combination of experimental and computational approaches has dramatically increased our knowledge of gene regulation in model organisms, relatively little is known about regulatory networks in other organisms. In this paper, we attempt to uncover the gene regulatory network in such lesser-studied organisms, using our knowledge of gene regulation in a well-studied organism.

This question can be placed in the larger context of comparative genomics. One goal of traditional comparative genomics is to infer the function of unknown proteins based on sequence homology to known proteins. More recently, the discovery of potential transcription factor binding sites by aligning gene regulatory regions across organisms has emerged as a second important goal of comparative genomics. Using sequence homology of the gene regulatory region in one organism to infer regulatory relations in another organism borrows from these two approaches, and can be seen as an emerging additional goal of comparative genomics.

To determine if the regulation of a gene in a given organism is conserved, we may attempt to search for potential binding sites of transcription factors that are known to regulate the homologous gene in a nearby organism. However, such a prediction is nontrivial for several reasons. First, since the number of experimentally known transcription factor binding sites is bounded, our statistical model of the binding motifs is necessarily limited and allows us to create only simplified models of the DNA motifs, such as position-weight matrices, or perhaps a first-order Markov model. Secondly, a transcription factor is not guaranteed to bind to each high-scoring DNA motif in vivo, as this may be affected by the presence of other DNA-binding proteins, the local bending of the DNA, or on cellular conditions. Third, it is often unclear if a transcription factor binding to a specific DNA site has a biological function, and if so, what the regulatory role might be. Finally, it is generally unknown if a transcription factor in a lesser-characterized organism recognizes the same DNA motifs as its homologous counterpart in a well-studied organism. Whereas aligning upstream sequences regions of homologous genes of different organisms has been successful in detecting transcription factor binding sites, it is unknown whether the conservation of binding motif extends to all transcription factors. At any rate, we expect some reduction in the accuracy of detecting transcription factor binding sites due to an imperfect conservation of the DNA binding specificity between two homologous transcription factors of different species.

While the prediction of gene regulatory networks based on a well-studied model organism is therefore a challenging task, the results that we may obtain from such an analysis are of tremendous importance. First, it allows us to assess whether our knowledge of model organisms represents biology in general, or if the gene regulatory network of each organism needs to be studied separately. Second, comparing the regulatory network in a model organism to those of lesser-known organisms will help us to identify regulatory subnetworks that do not occur in the model organisms, and that may therefore contain currently unknown biological mechanisms. Finally, it is of great medical importance to understand biological networks in disease-causing organisms. Whereas a considerable number of genome sequencing projects of such organisms have now been completed, only a fraction of their regulatory network is known. Examples of close relatives of the model organism *Bacillus subtilis*, which we consider in this paper, are *Bacillus anthracis*, which causes anthrax, *Staphylococcus aureus*, whose multiple-resistant form (MRSA) is a major source of hospital infections, and *Clostridium tetani*, the causative agent of tetanus.

Here, we focus on the regulatory network underlying initiation of sporulation in spore-forming bacteria. Gram-positive bacteria of the *Bacilli* and *Clostridia* genus have the capability of forming spores when environmental conditions become adverse. Spores, which are metabolically dormant, protect the bacterial DNA from environmental challenges such as heat, dryness, and UV radiation. As soon as the environmental conditions ameliorate, spores germinate to create the complete bacterium. Sporulation therefore helps these bacteria to survive prolonged periods of adverse environmental conditions.

The decision to sporulate has a profound effect on the survivability of a bacterium. By sporulating too soon, a bacterium inadvertently passes up additional rounds of replication, whereas failing to detect the need for sporulation may kill a bacterium altogether. *Bacillus subtilis* therefore contains an intricate regulatory network to initiate sporulation. This network connects the input of environmental sensors via two-component histidine kinases to the activation of the master regulator Spo0A, which regulates a large number of genes that are active in the initial stage of sporulation.

Transcription factors involved in sporulation as well as their regulated genes, as discovered experimentally, are collected in the DBTBS database of transcriptional regulation in *Bacillus subtilis* [1]. We use the information in this database to investigate whether the regulatory network of sporulation initiation in *Bacillus subtilis* is conserved in fourteen other fully-sequenced sporulating bacteria, ranging from the nearby *Bacillus licheniformis* to the more remote *Clostridia.*

Due to the difficulties of inferring the gene regulatory network in other organisms on the basis of sequence information, as noted above, we do not aim to identify each transcription factor binding site individually. Instead, given the combination of transcription factors that bind to the upstream region of a particular gene, we construct a joint scoring function that combines the scores of the individual binding sites. We assess the degree to which conservation is conserved by comparing the scores obtained for the homologous genes in the fourteen target organisms to a background distribution, obtained by calculating the scores for genes known not to be involved in regulation.

2 Method

2.1 Aligning known transcription factor binding sites

Our first task is to create a statistical model of the sequence motifs known to bind a particular transcription factor in *Bacillus subtilis*. Experimentally, transcription factor binding sites can be localized to short (about 20 nucleotide) DNA sequences, from which a conserved sequence motif (which is typically shorter) can be found by alignment. As a statistical model, we use a position-weight matrix approach [2]. The log-likelihood score of an alignment with $n\left(i,c\right)$ nucleotides c at position i can be written as

$$L = \sum_{i=1}^{m} \sum_{c=\{A,C,G,T\}} n\left(i,c\right) \log \frac{n\left(i,c\right)}{n}, \tag{1}$$

where m is the motif length and n is the number of sequences in the alignment; $p\left(i,c\right) = n\left(i,c\right)/n$ is the corresponding probability to find a nucleotide c at position i.

Some transcription factor binding sites consists of two sequence motifs separated by a gap for which the sequence is not conserved. In particular, the promoter sequence on the DNA, which is recognized by the σ (specificity) factor subunit of the RNA polymerase, consists of two motifs, located around 35 base-pairs and 10 base pairs upstream of the transcription start site. The two binding motifs are separated by a gap of variable length. To model such binding sites, we assume a flat probability distribution for gaps of $w_{\mathrm{min}}, \ldots, w_{\mathrm{max}}$ base pairs, and write the log-likelihood as

$$\begin{aligned} L = & \sum_{i=1}^{m_{\mathrm{left}}} \sum_{c=\{A,C,G,T\}} n_{\mathrm{left}}\left(i,c\right) \log \left(\frac{n_{\mathrm{left}}\left(i,c\right)}{n}\right) \\ & + \sum_{i=1}^{m_{\mathrm{right}}} \sum_{c=\{A,C,G,T\}} n_{\mathrm{right}}\left(i,c\right) \log \left(\frac{n_{\mathrm{right}}\left(i,c\right)}{n}\right) \\ & - n \log \left(w_{\mathrm{max}} - w_{\mathrm{min}} + 1\right). \end{aligned} \tag{2}$$

Sigma factor recognition sites can be found experimentally by determining the starting position of the mRNA molecule in a primer extension or S1 nuclease experiment.

We aligned the known transcription factor binding sites using BioProspector [3]. As repeated runs of BioProspector returned identical alignments, we feel confident that the optimal alignment solution was found. For the sigma factor recognition sites, we used BioProspector to align the 50 basepair regions upstream of the transcription start sites, trying all plausible values of w_{min} and w_{max} exhaustively and evaluating the alignment result using Equation (2). This allows us to find a statistically founded balance between the motif alignment scores and the allowable gap variation. We note that the allowable gaps we found are typically more restrictive than those identified previously [4].

To avoid the problem of overfitting, after alignment we added the pseudocounts $q_c\sqrt{n}$, where q_c is the background probability to find a nucleotide c. Hence, the position-weight matrix for transcription factor T can be written as

$$M_T\left(i,c\right) = \log\left(\frac{\left(n\left(i,c\right) + q_c\sqrt{n}\right)/\left(n+\sqrt{n}\right)}{q_c}\right) \tag{3}$$

2.2 Finding combinations of transcription factor binding sites

Given the position-weight matrix, we can search the upstream region of genes to find likely transcription factor binding sites. To reduce the effect of false-positives, we aim to find several predicted transcription factor binding sites located within a window of length w. The score for a potential transcription factor binding site s, after Bonferoni correction for multiple comparisons over the window of length w, is calculated from the position-weight matrix as

$$\text{score}\left(s, \text{transcription factor } T\right) = \sum_{i=1}^{m} M_T\left(i, s_i\right) - \log\left(w\right) \tag{4}$$

The joint score that a given set of transcription factors T bind to a potential regulatory region of length w is the calculated as

$$\text{score(sequence of length } w) = \sum_{T} \sum_{\text{all subsequences } s \text{ of length } m_T} R\left(\text{score}\ \left(s,T\right)\right), \tag{5}$$

where R is the ramp function defined by

$$R(x) = \begin{cases} x \text{ if } x > 0; \\ 0 \text{ otherwise} \end{cases} \tag{6}$$

Effectively, we include every potential transcription factor binding site in the score function (5) with a positive log-likelihood score after Bonferoni correction. To determine if the regulation of a gene is conserved in a target organism, we slide a sequence window of length w along its upstream sequence region and calculate the score function (5), including only those transcription factors T in the summation that are known to regulate the gene in the source organism *Bacillus subtilis.*

To facilitate their interpretation, we normalize the joint scores as defined by Equation (5) by dividing them by the root mean square of the scores found for 2264 *Bacillus subtilis* genes that are known not to be involved in sporulation, based on their functional annotation in the SubtiList database [5, 6]. As these genes are unlikely to be regulated by the sporulation-specific transcription factors, their scores can serve as a background model.

3 Results

We search the upstream DNA regions of genes whose homologs in *Bacillus subtilis* are regulated by Spo0A, the master regulator of sporulation initiation. In addition to Spo0A, these genes are regulated by various combinitions of the transcription factors AbrB, AhrC, CcpA, Hpr, SinR, and SpoVT, and the sigma factors SigA, SigD, SigF, SigG, SigH, and SigX. As no binding site information is available for SpoVT, we excluded this transcription factor from our analysis. For each gene, we use Equation (5) to assess the overall resemblance of the upstream regulatory DNA sequences to their *Bacillus subtilis* counterpart.

Table 1 shows the prediction result for *Bacillus subtilis* itself, two strains of *B. licheniformis*, three strains of *B. anthracis*, three strains of *B. cereus*, one strain each of *B. thuringiensis*, *B. clausii*, *B. halodurans*, and three *Clostridia*. Gene regulation appears to be well-conserved in *Bacillus licheniformis*, for which high scores were found for nearly all genes under consideration. A lesser degree of conservation was found for the *Bacillus anthracis*, *cereus*, and *thuringiensis* strains, for which high-scoring regulatory regions were found for only half of the genes. For the *Clostridia*, regulation was usually not conserved, in particular for *Clostridium perfringens*.

We note that the degree of conservation varies strongly between genes. For the *spo0A* gene, which encodes the master regulator of sporulation initiation, regulation appears to be conserved in nearly all species, including two of the *Clostridia*. Similarly, for the *spoIIAA-spoIIAB-sigF* operon, we find a highly conserved regulatory region. Interestingly, the products of these genes play key roles in the regulation of sporulation initiation. SpoIIAB is an anti-sigma factor that inhibits SigF by binding to it; SpoIIAA is an anti-anti-sigma factor that binds to SpoIIAB, thereby releasing the inhibition of SigF. SigF is the first sporulation-specific sigma factor to be activated in sporulation, representing a major step in the initiation of sporulation. Similarly, regulation is conserved in most species for *abrB*, whose product prevents the expression of sporulation related genes until the start of sporulation, as well as for *spo0F*, a two-component response regulator involved in sporulation initiation.

Of the genes whose regulation is less conserved, *dltABCDE*, *argCJBD-carAB-argF*, and *rbsRKDACB* do not play central roles in the sporulation network. The products of the *argCJBD-carAB-argF* and *dltABCDE* operon are enzymes involved in arginine and lipotheichoic acid biosynthesis, respectively; the *rbsRKDACB* genes code for proteins involved in ribose transport and their regulator. SpoIIE, however, a serine phosphatase acting on SpoIIAA, is importation in the sporulation initiation network. The regulatory region of *spoIIE* consists of several weak Spo0A binding sites, none of which are strong enough to overcome the Bonferoni correction. Hence, we are unable to detect the regulatory region even in *Bacillus subtilis* itself. Similarly, for the *spoIIGA-sigE-sigG* operon, also important for the initiation of regulation, only one of the weak transcription factor binding sites can be detected in *Bacillus subtilis*.

For *kinC* and *kinA*, whose products act as two-component sensor histidine kinases in the phosphorylation pathway of sporulation initiation, the main dif-

Table 1. Conservation of regulation of sporulation initiation in *Bacilli* and *Clostridia*. The listed numbers are the scores calculated from Equation (5), divided by the root mean square of the scores calculated for genes known not to be involved in regulation. To calculate the score, we include include those transcription factors that are known to regulate the gene in *Bacillus subtilis*. A dash indicates that no orthologous gene could be identified in the genome.

Organism	*abrB*	*argC*[a]	*dltA*[b]	*kinA*	*kinC*	*rbsR*[c]	*sinI*[d]	*spo0A*	*spo0F*	*spoIIAA*[e]	*spoIIE*	*spoIIGA*[f]
B. subtilis	5.1	2.9	8.2	5.0	0.2	2.7	1.8	4.1	3.0	4.0	0.0	0.4
B. licheniformis Novozymes Biotech	2.3	2.4	5.1	5.3	0.0	0.9	0.1	3.8	2.3	4.6	0.0	0.4
B. licheniformis Göttingen	2.3	2.4	5.1	5.3	0.0	0.2	0.1	3.8	2.3	4.6	0.0	0.4
B. anthracis str. ‘Ames Ancestor’	2.3	2.0	0.0	0.0	0.3	0.2	1.9	4.8	5.7	4.0	0.0	0.0
B. anthracis str. ‘Ames’	2.3	2.0	0.0	0.0	0.3	0.2	1.9	4.8	5.7	4.0	0.0	0.0
B. anthracis str. Sterne	2.3	2.0	0.0	1.5	0.3	0.2	1.9	3.5	5.7	4.0	0.0	0.0
B. cereus ATCC 10987	2.3	2.0	0.0	0.0	3.1	0.2	1.9	3.7	4.4	3.7	0.0	0.0
B. cereus ATCC 14579	2.3	0.3	0.0	0.7	0.0	0.2	0.0	4.8	5.6	4.0	0.0	0.0
B. cereus ZK	1.2	2.0	0.0	0.0	1.4	0.0	-	3.5	5.2	4.0	0.0	0.1
B. thuringiensis	0.0	2.0	0.0	0.0	0.9	0.4	-	3.5	5.8	4.0	0.0	0.0
B. clausii	5.5	1.2	-	0.0	0.0	0.0	-	3.8	3.3	2.2	2.8	0.8
B. halodurans	6.5	0.5	-	0.0	0.8	1.0	0.0	3.4	4.2	1.7	0.0	0.4
C. acetobutylicum	0.0	2.4	-	1.0	0.0	0.5	-	4.0	-	3.5	0.2	0.0
C. perfringens	4.1	-	-	0.0	1.6	0.5	-	1.4	-	0.1	0.0	1.5
C. tetani	0.0	-	-	2.4	0.0	1.1	-	4.3	-	2.4	0.0	0.0

[a] *argCJBD-carAB-argF* operon

[b] *dltABCDE* operon

[c] *rbsRKDACB* operon

[d] *sinIR* operon

[e] *spoIIAA-spoIIAB-sigF* operon

[f] *spoIIGA-sigE-sigG* operon

ficulty lies in the determination of their homologs in other species. At least five histidine kinases, with overlapping roles, participate in the initiation of sporulation in *Bacillus subtilis*. As no clear one-to-one relation exists between histidine kinases in different organisms, except in the most nearby organisms, we do not expect to be able to find a strong conservation of their regulatory regions.

4 Discussion

The example of the regulation of sporulation initiation demonstrates some of the potentials and pitfalls of comparative systems biology.

Our analysis of the regulatory regions of genes involved in sporulation initiation reveals that the core of the network appears to be conserved between organisms, whereas more peripheral parts of the network are poorly conserved. In addition to the difficulty in recognizing regulatory regions, the identification of orthologous genes in different can be non-trivial. We found this especially to be the case for the histidine kinases, which are important in phosphorylation signalling pathways in the initiation of sporulation. This suggests that a comprehensive approach, in which signalling pathways and gene regulatory relations are reconstructed simultaneously, may give a clearer view of the conservation of regulatory networks.

As our scoring function (Equation (5)) takes into account the contribution of several transcription factors, it is more powerful in detecting potential regulatory regions and their conservation. However, a more detailed analysis of regulatory DNA, in which individual transcription factor binding sites are identified (if feasible) is preferable. We feel that currently, our ability to detect transcription factor binding sites is not sufficiently powerful for such a prediction. In addition to improving the accuracy of predicting transcription factor binding sites, another challenge is to determine which of the transcription factor binding sites fulfill a biological function, and if so, what that function is.

References

1. Makita, Y., Nakao, M., Ogasawara, N., and Nakai, K.: DBTBS: Database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics. Nucleic Acids Res., 32 (2004) D75–77.
2. Durbin, R., Eddy, S.R., Krogh, A., and Mitchison, G.: Biological sequence analysis. Cambridge University Press, Cambridge, UK (1998).
3. Liu, X., Brutlag, D.L., and Liu, J.S.: BioProspector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac. Symp. Biocomput., **6** (2001) 127–138.
4. Sonenshein, A.L., Hoch, J.A., and Losick, R.: *Bacillus subtilis* and its closest relatives: from genes to cells. ASM Press, Washington, DC (2001).
5. Moszer, I., Glaser, P., and Danchin, A.: SubtiList: a relational database for the *Bacillus subtilis* genome. Microbiology, **141** (1995) 261–268,
6. Moszer, I.: The complete genome of *Bacillus subtilis*: From sequence annotation to data management and analysis. FEBS Letters **430** (1998) 28–36.