BacillusRegNet: A transcriptional regulation database and analysis platform for *Bacillus* species

Goksel Misirli[1], Jennifer Hallinan[1], Richard Röttger[2], Jan Baumbach[2], Anil Wipat[1,*]

[1]School of Computing Science, and Centre for Synthetic Biology and Bioexploitation, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK

[2]Department of Mathematics and Computer Science, University of Southern Denmark, Odense, DK-5230, Denmark

Summary

As high-throughput technologies become cheaper and easier to use, raw sequence data and corresponding annotations for many organisms are becoming available. However, sequence data alone is not sufficient to explain the biological behaviour of organisms, which arises largely from complex molecular interactions. There is a need to develop new platform technologies that can be applied to the investigation of whole-genome datasets in an efficient and cost-effective manner. One such approach is the transfer of existing knowledge from well-studied organisms to closely-related organisms. In this paper, we describe a system, BacillusRegNet, for the use of a model organism, *Bacillus subtilis*, to infer genome-wide regulatory networks in less well-studied close relatives. The putative transcription factors, their binding sequences and predicted promoter sequences along with annotations are available from the associated BacillusRegNet website (http://bacillus.ncl.ac.uk).

1 Introduction

As Next Generation Sequencing technologies become cheaper and easier to use, the number of available genome sequences is increasing exponentially. Large amounts of data about raw sequences and their annotations are publicly available in biological databases for many organisms. At the time of writing, the Genomes OnLine Database[1] lists 12,856 complete and 26,308 incomplete genomes [1]. However, in order to understand the biology of these organisms, interactions between different biological molecules must be understood in detail.

In many cases, transcriptional-level interactions are easier to understand, infer and engineer than complex protein-protein interactions. Transcriptional regulatory networks can be used to convert environmental signals into cell-level biological signals and to regulate metabolic pathways. Although recent developments in high-throughput technologies such as transcriptome analysis and microarray experiments provide the data needed for biologists to investigate the regulatory relationships, carrying out these experiments for every new species is neither cost- nor time-effective. There is a large amount of information already available for model organisms such as *Bacillus subtilis* and *Escherichia coli*. This information can be systematically analysed and used to infer information about close relatives.

B. subtilis and its relatives are widely used in the biotechnology industry. *B. subtilis* is a Gram-positive, non-pathogenic, model organism and is generally considered to be safe [2].

Copyright 2014 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

[*] To whom correspondence should be addressed. Email: anil.wipat@newcastle.ac.uk

[1] http://genomesonline.org/cgi-bin/GOLD/index.cgi

doi:10.2390/biecoll-jib-2014-244

The organism inhabits the soil and can develop symbiotic relationships with plants [3]. *B. subtilis* is well studied, and its amenability to genetic manipulation makes it ideal for laboratory studies [4-6]. Other *Bacillus* species are also industrially important. *B. amyloliquefaciens* is known to promote plant growth [7], and *B. megaterium* is used to produce vitamin B12 [8]. *B. licheniformis* is also used in the production of antibiotics and enzymes [9]. There are also related species such as *Geobacillus* spp., some of which were initially classified under the genus *Bacillus* [10]. These species can be thermophilic, and hence can be used in high temperature environments, for example to metabolise hydrocarbons in oil fields [11]. In order to optimise these functions, and even create novel behaviours, the transcriptional control systems of these species must be well understood.

Although the genomes of many non-model organisms have been sequenced, detailed information about the regulatory networks of these organisms is not typically available. There are over 80 sequenced *Bacillus* strains in the NCBI genome database[2]. The NCBI's taxonomy browser[3] lists over 20,550 taxonomy terms for different *Bacillus* strains. However, complete detailed information about transcription factors (TFs), their binding sequences and promoters is not available for non-model *Bacillus* species.

Comparative genomic approaches can be used to infer the transcriptional regulatory networks of non-model *Bacillus* species. Taxonomic distance has been shown to correlate with measures of similarity between gene regulatory networks and, hence, the comparison of genomes in order to identify conserved genes has been used to predict gene regulatory networks [12]. This information is useful to elucidate the relationships between TFs and target genes in poorly-studied organisms. However, further details such as the sequences of TF binding sites (TFBSs) are needed in order to facilitate the engineering of regulatory networks. Moreover, the availability of TFBS sequences may aid in the inference of the regulatory networks [13, 14], increasing their quality [12, 15]. Searching for these sequences can be facilitated by the use of position weight matrices (PWMs) [16, 17], which represent TF binding motifs, and can be used as input to motif finding tools such as PoSSumSearch [18] and MAST [19]. Applications such as RegNet [20], FITBAR [21] and RegPredict [22] provide platforms for TFBS predictions, using existing methods and tools, and allowing Web access to results.

1.1 The RegNet system

RegNet is a system that reconstructs prokaryotic transcriptional regulatory networks on a genome-wide scale by combining TFBS searches with detection of orthology between genes in different species. This approach improves the construction of PWMs by using data from the reverse strand of binding sequences. Binding motifs are also optimised by modifying the predicted binding sequences using a sliding window [20]. Predictions are associated with *p*-values indicating their statistical significance.

The RegNet system uses model organisms to predict the regulatory networks of their close relatives, and was initially produced for the *Corynebacteria* [20]. The experimentally verified gene regulatory networks of *Corynebacterium glutamicum* ATCC 13032 were used to predict TFs and their binding sequences in other *Corynebacterium* species. The RegNet system was then extended to include *E. coli* [23] and the Mycobacteria [24]. The results were stored in the publicly-accessible databases CoryneRegNet [20, 23-26] and MycoRegNet [27].

[2] ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/ (accessed 30/01/2014)

[3] http://www.ncbi.nlm.nih.gov/taxonomy/?term=txid1386[Subtree] (accessed 30/01/2014)

Copyright 2014 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

RegNet integrates a number of different types of data, such as the nucleotide and amino acid sequences of coding sequences (CDSs) and proteins, operon annotations and known gene regulatory relationships (Figure 1). The system initially detects orthology between genes in the model organism and those in the target organisms, and searches upstream of these genes to find TF binding sequences, using information from the model organism. The list of predicted interactions is then imported into a database. All-versus-all BLAST is then run to detect sequence similarities. The BLAST results are used as an input to a protein clustering algorithm to detect protein homologies [24, 25]. If the binding sites are conserved and their regulated genes are homologous, the predicted interactions are more reliable [20]. Therefore, interactions that have both homologous TFs and target gene pairs are searched for in terms of binding sequences in the relevant species.

Binding sequences identified are used to construct PWMs, which are then searched for on a genome-wide scale in target organisms using the PoSSumSearch tool [18]. A list of binding sequences is produced for conserved TFs and target genes between the model organisms being used and their closely related species. It is assumed that the role of an interaction is also conserved, irrespective of whether the regulation is positive or negative [27].

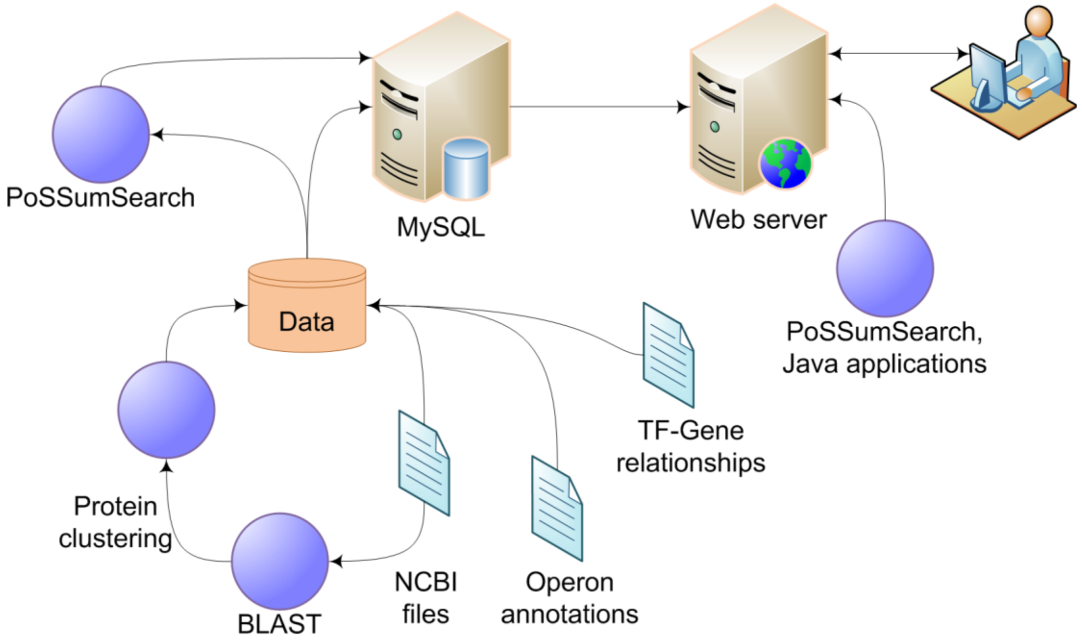


Figure 1: The components of the RegNet system. NCBI files, known gene regulatory relationships, and operon annotations are integrated. Protein homologies are inferred using BLAST and a protein clustering algorithm. PoSSumSearch is used to search for binding sequences. The results can then be accessed via a Web interface. Users can also choose custom parameters to search for these sequences and can visualise the results via this interface.

We extended the approach used in RegNet to *B. subtilis* in order to construct genome-wide gene regulatory networks for related *Bacillus* species. There is a large amount of genetic and metabolic data available for *B. subtilis*, which can thus be used as a template organism in RegNet in order to transfer experimentally obtained information about its gene regulatory networks to its close relatives.

2 BacillusRegNet construction

Data about TFs and their binding sequences from DBTBS [28], and the latest annotations for these TFs from BacilluScope [2] were integrated into the BacillOndex knowledgebase [29,

Copyright 2014 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

doi:10.2390/biecoll-jib-2014-244

30]. BacillOndex was constructed using Ondex [31], a graph-based data integration tool, which can also be used for querying or visualising the data. A subset of this dataset containing the details of the gene regulatory networks of *B. subtilis* was used as an input into the RegNet system. Data about the gene regulatory relationships was created as a tab-delimited file. The file contains a row for each TF and its regulated CDS. Data from BacillOndex was mapped to RegNet's format, as shown in Table 1. TF family names and accessions representing the locus tags were used for the mapping. If a CDS concept in the BacillOndex dataset has positive or negative autoregulation recorded as an attribute, the corresponding 'Auto' entry in the RegNet system is populated with '+' and '-' respectively. The CDS gene module is taken from the biological role classification attribute of the corresponding CDS concept. The characters 'A' and 'R' are used to represent the role of the TF as activator or repressor. The PubMed identifiers (PMIDs) and binding sequences were recorded as semi-colon separated lists. For TFs that are sigma factors, the binding sequences represent the core promoter sequences to which the RNA polymerases (RNAPs) bind. These TFs are identified with the 'Is CDS Sigma factor' field. Since the RegNet system only deals with proteins, genes that encode non protein coding RNAs were excluded.

Table 1: List of gene regulation fields required for the RegNet system and the mapping from the BacillOndex dataset.

RegNet field	Values mapped from BacillOndex
CDS	CDS accession
CDS gene name	CDS name
CDS gene module	Role classification
TF Family	TF family name such as ArsR, GntR and sigma factor
Auto	`Empty` : if not auto regulatory `-` : if negative auto regulatory `+` : if positive auto regulatory
Role	Role of the binding `Empty` : If not known `A` : If activator `R` : If repressor
Target gene	Target CDS accession
Target gene name	Target CDS name
Target gene module	Role classification of the target CDS
Motif known	`known` : If the binding motif is known `-` : If the binding motif is not known
Evidence	`Experimental`: To indicate that experimental information is known
PubMed IDs	Semi colon separated list of PMIDs
Binding motif	Semi colon separated list of TF binding sequences
Is CDS Sigma factor	`+` : If the CDS is encoding for a sigma factor

The resulting file included the binding sequences for transcription factors from *B. subtilis*. These sequences may include additional upstream and downstream sequences that are not part of the actual binding sites. As a result, the entire sequence may not be conserved. Although most of the binding sequences have annotations indicating the actual binding sequences, not all are annotated. However, PWMs, which are used to search for TFBSs in target organisms, require all of the binding sequences for a particular TF to be the same length. Therefore, binding sites must be aligned and the additional bases from upstream and downstream of the binding sites excluded. This process was carried out by using the MEME tool [19].

Copyright 2014 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

The parameters of MEME were adjusted to find a binding motif for each TF. Each TF binding sequence was required to possess the presence of binding motif, which was also searched for in reverse and complementary sequences. The maximum length of a binding motif was set to the smallest observed length of the binding sequence, and the minimum length was set to 90% of the smallest sequence length for the TF. MEME was run for every TF having more than one binding sequence available. The resulting sequences were used as the new binding sequences in genome-wide searches (Figure 2).

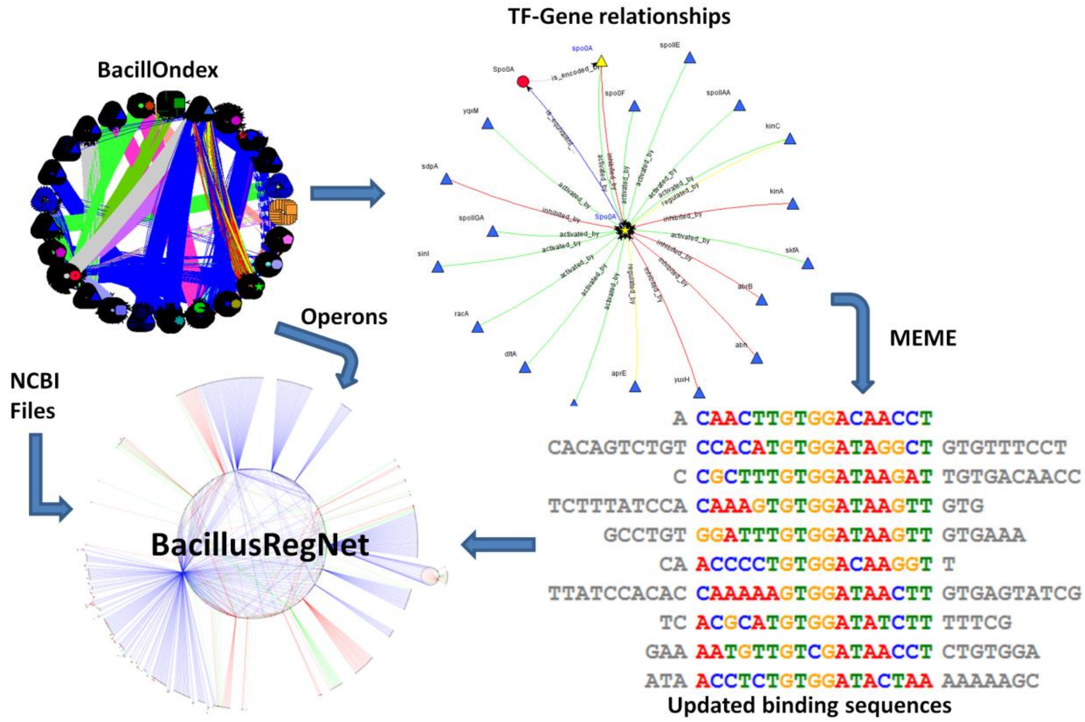

TF-Gene relationships
BacillOndex
Operons
NCBI Files
BacillusRegNet
MEME
A CAACTTGTGGACAACCT
CACAGTCTGT CCACATGTGGATAGGCT GTGTTTCCT
C CGCTTTGTGGATAAGAT TGTGACAACC
TCTTTATCCA CAAAGTGTGGATAAGTT GTG
GCCTGT GGATTTGTGGATAAGTT GTGAAA
CA ACCCCTGTGGACAAGGT T
TTATCCACAC CAAAAAGTGGATAACTT GTGAGTATCG
TC ACGCATGTGGATATCTT TTTCG
GAA AATGTTGTCGATAACCT CTGTGGA
ATA ACCTCTGTGGATACTAA AAAAAGC
Updated binding sequences

Figure 2: The binding sequences for each TF from the BacillOndex dataset were adjusted to have the same length using the MEME tool in order to create PWMs. These matrices are used to search for binding sequences in other organisms.

For example, in the system, eight binding sequences exist for the RocR TF. These sequences range from 21 to 29 bp in length (Figure 3). The motif was searched for with 19 and 21 as the minimum and maximum length respectively. The motif was constructed using the 19 bp sequences (Figure 4).

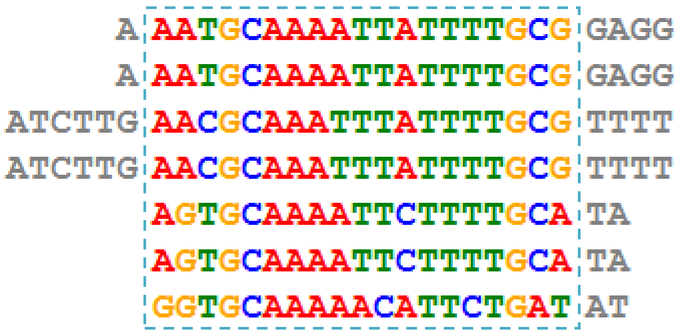

A AATGCAAAATTATTTTGCG GAGG
A AATGCAAAATTATTTTGCG GAGG
ATCTTG AACGCAAATTTATTTTGCG TTTT
ATCTTG AACGCAAATTTATTTTGCG TTTT
AGTGCAAAATTCTTTTGCA TA
AGTGCAAAATTCTTTTGCA TA
GGTGCAAAAACATTCTGAT AT

Figure 3: The binding motif for the RocR TF. The coloured bases, inside the rectangle, show the sequences that form the binding motif.

Copyright 2014 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

FASTA files for the nucleotide sequences and GenBank files for all organisms were downloaded from NCBI[4]. For each non-model organism operon predictions were downloaded from Microbes Online [32] and converted into the RegNet format, using a utility provided in RegNet. For *B. subtilis*, this information was extracted from the BacillOndex dataset [30]. The final, integrated dataset with its associated analytical tools comprises BacillusRegNet.

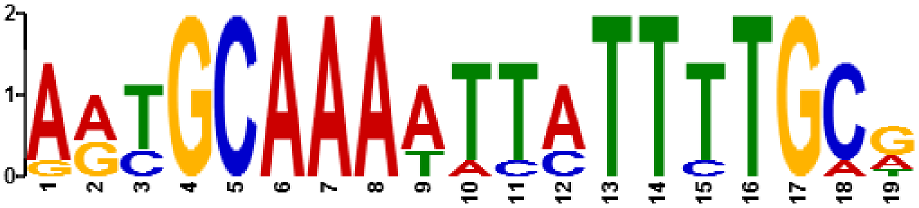


Figure 4: The sequence logo constructed for the binding motif of the RocR TF.

3 BacillusRegNet: A platform for the analysis and transfer of *Bacillus* gene regulatory networks

We used BacillusRegNet[5] to infer the gene regulatory networks of *B. subtilis* 168 and 15 of its relatives including 13 *Bacillus* and two *Geobacillus* species. The list includes strains from *B. amyloliquefaciens* (FZB42), *B. licheniformis* (ATCC 14580), *B. pumilus* (SAFR-032), *B. megaterium* (DSM 319), *B. halodurans* (C-125), *B. anthracis* (A0248 and Sterne)¸ *B. clausii* (KSM-K16), *B. thuringiensis* (Al Hakam), *B. cytotoxicus* (NVH 391-98), *B. weihenstephanensis* (KBAB4), *B. cereus* (B4263), *B. tusciae* (DSM 2912), *G. kaustophilus* (HTA426) and *G. thermodenitrificans* (NG80-2). Some of these organisms host plasmids, the details of which are also available in BacillusRegNet. The system infers genome-wide gene regulatory networks for these non-model organisms using the well-understood TFs, and their target genes and binding sequences from *B. subtilis* 168. The system includes predictions for the nucleotide sequences of TF binding sites and promoters in target organisms, and homology information about CDSs and proteins.

3.1 Genome-wide construction of gene regulatory networks

BacillusRegNet provides two databases, Experimental and Predicted, each with a Web user interface (Table 3). The Experimental database contains information about experimentally-confirmed regulatory relationships from DBTBS, and gene annotations from BacilluScope, together with information derived from GenBank files and nucleotide sequences. The experimental data include 69,388 genes and proteins for *Bacillus* species, and information about the gene regulatory networks of *B. subtilis* 168 for 140 TFs, 1,148 binding sequences and 1,250 regulatory relationships (Table 2). The Predicted database stores predictions based on the experimental gene regulatory relationships, in addition to containing all the data from the database of experimental data. The system was used to predict an additional 696 TFs, 7,856 binding sequences and 14,540 regulatory relationships for 15 non-model organisms (Table 3). For all organisms in the system, the number of activatory and repressional interactions are 11,239 and 4,551 respectively.

[4] ftp://ftp.ncbi.nih.gov/genomes/Bacteria

[5] http://bacillus.ncl.ac.uk

Copyright 2014 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

Table 2: Summary of the databases containing experimental and predicted data. ‘Experimental’ data include information about the gene regulatory networks of *B. subtilis* 168 only. The ‘predicted’ database includes everything from the experimental data and the predictions for 15 other non-model organisms in the system.

	Experimental	**Predicted**
Genes	69388	69388
Proteins	69388	69388
Regulations	1250	15790
Regulators	140	836
Regulated genes	787	9349
Binding motifs	1148	9004
Position weight matrices	91	784
Protein clusters	7081	7081
Genomes	26	26

Table 3: Number of predicted regulators, binding sequences, regulatory relationships and regulated genes are listed for each organism. Genome sizes for the organisms were taken from the NCBI's GenBank database.

	Proteins	**Regulatory relationships**	**Regulators**	**Binding sequences**	**Regulated genes**	**Genome size (Mb)**
B. subtilis 168	4,175	1,250	140	1,148	787	4,215
B. amyloliquefaciens (FZB42)	3,693	1,611	75	875	870	3,918
B. halodurans (C-125)	4,065	845	43	443	498	4,202
B. anthracis (A0248)	5,040	862	43	466	528	5,227
B. anthracis (Sterne)	5,289	945	45	559	494	5,228
B. licheniformis (ATCC 14580)	4,173	989	72	989	567	4,222
B. clausii (KSM-K16)	4,096	983	43	455	566	4,303
B. thuringiensis (Al Hakam)	4,736	1,043	47	499	623	5,257
B. cytotoxicus (NVH 391-98)	3,833	821	36	359	496	4,087
B. pumilus (SAFR-032)	3,679	1,377	61	668	741	3,704
B. weihenstephanensis (KBAB4)	5,155	1,056	46	502	626	5,262
B. cereus (B4264)	5,398	901	43	517	529	5,419
B. tusciae (DSM 2912)	3,150	284	14	121	249	3,384
B. megaterium (DSM 319)	5,100	1,287	52	768	732	5,097
G. kaustophilus (HTA426)	3,497	787	37	378	492	3,544
G. thermodenitrificans (NG80-2)	3,392	748	38	326	480	3,550

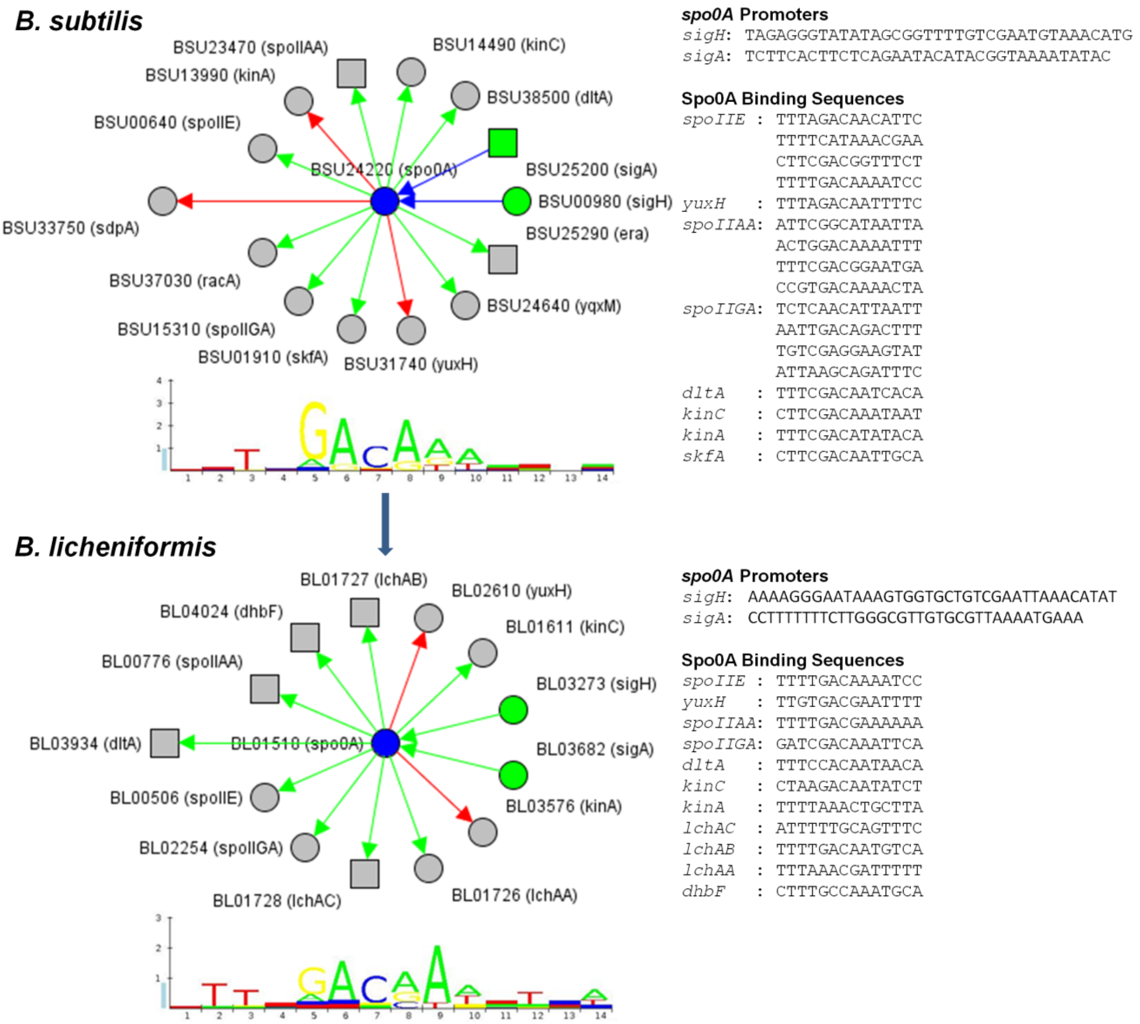
3.2 Analysis of the gene regulatory networks using BacillusRegNet

BacillusRegNet provides both text- and graph-based data visualisation using HTML and GraphVis respectively. The data can be searched using gene and protein identifiers. Details for each gene include information about the protein product, TFs that regulate the gene, whether the gene encodes a TF, genes that are regulated by the TF, homologous genes and proteins, gene attributes, a PWM, and the sequence logo used to depict the binding motif.

Figure 5 shows an example of a set of gene regulatory relationships predicted using BacillusRegNet. In the figure, the transcriptional network of the *spo0A* gene is visualised using GraphVis for *B. subtilis* 168 and the target organism *B. licheniformis* (ATCC 14580). The inhibition of *kinA* and *yuxH* by Spo0A, and the activation on *spoIIE, spoIIAA, spoIIGA*, *dltA* and *kinC* are predicted in the regulatory network of *B. licheniformis* (ATCC 14580).

Copyright 2014 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

However, the inhibition of *sdpA* and activation of *skfA* by Spo0A was not predicted since there are no homologues of these genes and their encoded proteins in the target organism *B. licheniformis* (ATCC 14580). Although there are homologues of *era* and *yqxM* in the target organism, the positive regulation of these genes by Spo0A could not be identified. The system was able to predict the *sigA* and *sigH* promoters of the *spo0A* gene. Additionally, activation relationships for *lchAA*, *lchAB*, *lchAC* and *dhbF* were predicted. The sequence logo constructed from predictions for *B. licheniformis* (ATCC 14580) is similar to that of *B. subtilis* 168.



***spo0A* Promoters**

```
sigH: TAGAGGGTATATAGCGGTTTTGTCGAATGTAAACATG
sigA: TCTTCACTTCTCAGAATACATACGGTAAAATATAC
```

Spo0A Binding Sequences

```
spoIIE : TTTAGACAACATTC
         TTTTCATAAACGAA
         CTTCGACGGTTTCT
         TTTTGACAAAATCC
yuxH   : TTTAGACAATTTTC
spoIIAA: ATTCGGCATAATTA
         ACTGGACAAAATTT
         TTTCGACGGAATGA
         CCGTGACAAAACTA
spoIIGA: TCTCAACATTAATT
         AATTGACAGACTTT
         TGTCGAGGAAGTAT
         ATTAAGCAGATTTC
dltA   : TTTCGACAATCACA
kinC   : CTTCGACAAATAAT
kinA   : TTTCGACATATACA
skfA   : CTTCGACAATTGCA
```

***spo0A* Promoters**

```
sigH: AAAAGGGAATAAAGTGGTGCTGTCGAATTAAACATAT
sigA: CCTTTTTTTCTTGGGCGTTGTGCGTTAAAATGAAA
```

Spo0A Binding Sequences

```
spoIIE : TTTTGACAAAATCC
yuxH   : TTGTGACGAATTTT
spoIIAA: TTTTGACGAAAAAA
spoIIGA: GATCGACAAATTCA
dltA   : TTTCCACAATAACA
kinC   : CTAAGACAATATCT
kinA   : TTTTAAACTGCTTA
lchAC  : ATTTTTGCAGTTTC
lchAB  : TTTTGACAATGTCA
lchAA  : TTTAAACGATTTTT
dhbF   : CTTTGCCAAATGCA
```

Figure 5: The transcriptional network of the *spo0A* gene for *B. subtilis* 168 and *B. licheniformis* (ATCC 14580). Red and green lines represent the inhibition and activation interactions, respectively. Similarly, red and green shapes represent the activators and inhibitors. Blue lines for the model organism show the sigma factor regulatory relationships. Squares and circles are used to represent whether or not the genes are part of operons and preceeded by TF binding sequences respectively. Sequence logos that depict the binding motifs are shown below each network. The sequences on the right-hand side are the known and predicted promoter and TF binding sequences for *B. subtilis* 168 and *B. licheniformis* (ATCC 14580) in BacillusRegNet respectively.

TF binding sequence predictions can also be performed manually, allowing user-defined cut-off values for the predictions of gene regulatory relationships to be input. Binding sequences can be searched for in two ways. In the first option, upstream sequences of all genes for a target organism are searched to find the genes that are regulated by the TF encoded by the queried gene. In the second option, TFs that regulate the queried gene are listed. In both cases,

Copyright 2014 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

the upstream sequences of the genes are searched. After running a transcription factor binding site (TFBS) search, both the source genes and the target genes can be visualised using GraphVis, with the relationships added for homologue proteins between the target and the source organisms. Additionally, TFBS predictions can also be achieved by submitting the binding sequences to be searched for as FASTA files through the TFBSScan section of the BacillusRegNet website.

BacillusRegNet also contains data about *B. subtilis* 168 core promoters that include the binding sites for RNAPs. Therefore, the system also predicts promoters in closely related *Bacillus* species. Figure 6 shows the SigA sequence logos for *B. subtilis* 168 and *B. amyloliquefaciens* (FZB42) aligned with the consensus sequence TTGACA-N$_{14}$-tgnTATAAT [33]. Compared to 317 experimentally known *sigA* promoters in *B. subtilis* 168, 574 *sigA* promoters were predicted for *B. amyloliquefaciens* (FZB42). As can be seen, the predicted sequence logo is also similar to the consensus sequence. The *sigA* gene is known as *rpoD* in *B. amyloliquefaciens* (FZB42). This information is available in the list of *B. subtilis* 168 *sigA* homologues.

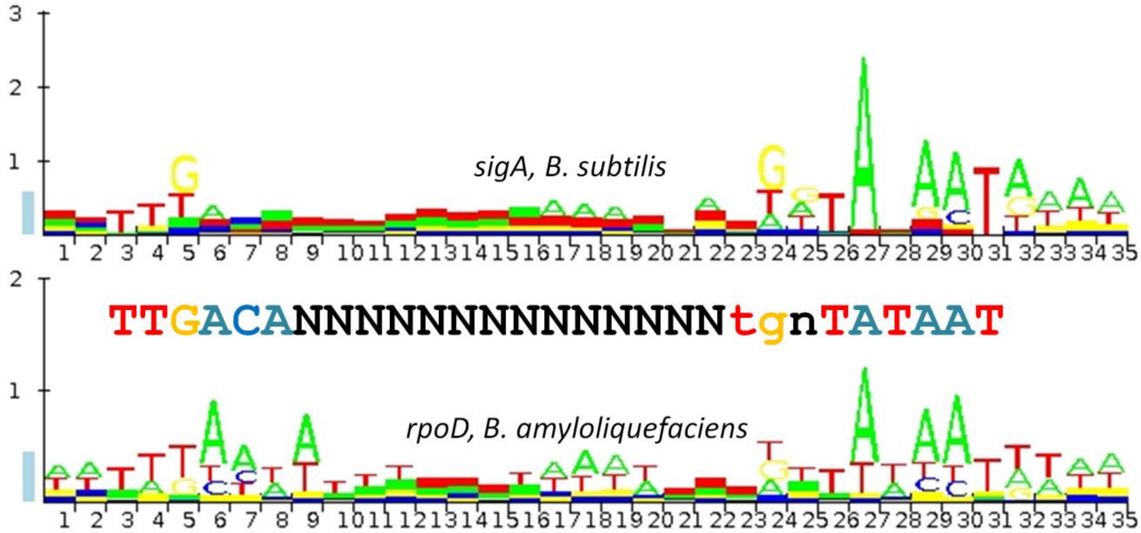


Figure 6: The sequence logos of *sigA* promoters for *B. subtilis* 168 and *B. amyloliquefaciens* (FZB42). The logo at the top is drawn using 317 core promoter sequences that includes the binding sequences for RNAPs from *B. subtilis* 168. The logo at the bottom was constructed from 574 predicted *B. amyloliquefaciens* (FZB42) core promoters. The middle sequence shows the consensus *sigA* promoter sequence, aligned with two sequence logos.

Genome-wide statistics can also be accessed from the website. The statistics provided include the distribution of binding sites from the genes' start locations, quantities of regulators and regulation types, and the distribution of the number of co-regulating TFs. ATGC content for the genome, and coding and non-coding regions are also available for each organism in the system. Figure 7 shows the distribution of TFBS distances from the start of genes for *G. kaustophilus* (HTA426). The number of repressors and activators predicted for G. *kaustophilus* (HTA426) is 17 and 15. Additional five TFs have dual roles. However, compared to 552 activation relationships, only 235 repression relationships were predicted. These activators and repressor sites tend to be between 0 and +100 relative to the gene start locations.

Copyright 2014 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

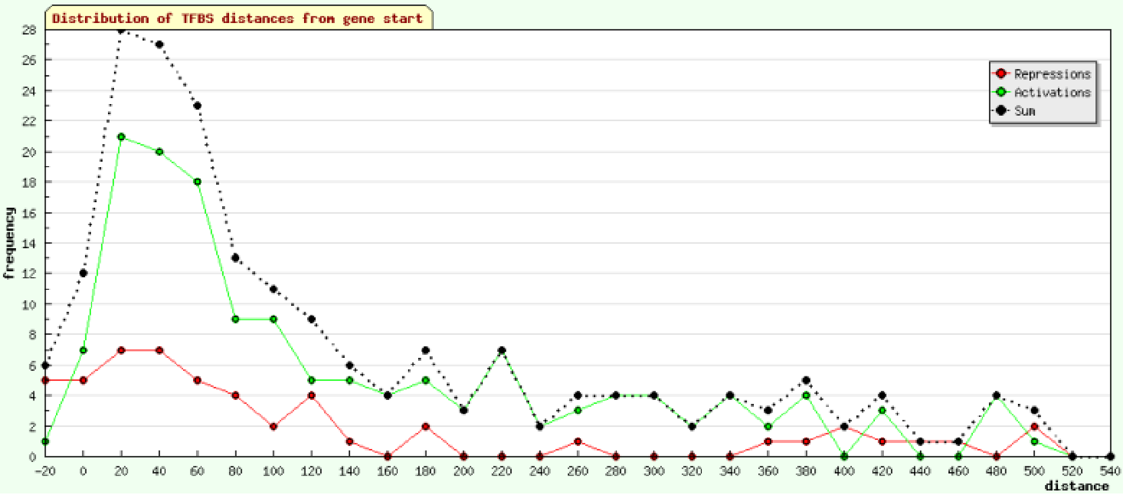


Figure 7: An example of the distribution of TF binding site distances from gene start for *G. kaustophilus* (HTA426). Red and green lines show the distribution of repressions and activations respectively. The distribution for all regulations is represented with the dashed black lines.

4 Discussion

By developing the BacillusRegNet system described in this paper, we aim to extend the gene regulatory information available to researchers interested in working with organisms related to *B. subtilis* 168. In turn, use of the system should help increase our understanding of gene regulation in closely-related *Bacillus* species by providing putative gene regulatory networks for a range of these organisms. These predicted networks can then be used to inform experimental design in non-model *Bacillus* species [34].

The BacillusRegNet approach can be applied to any related species using the scripts built into the RegNet system. Currently there are, in the BacillusRegNet database, 16 species and 836 TFs, 696 of which have been predicted to exist as orthologs in the 15 non-model organisms. For these non-model organisms 7,856 target binding sequences, including promoters, were predicted. The number of TFs predicted for an organism varies due to several factors.

Firstly, the size of the total regulatory network varies between organisms. For example, the ecological niche occupied by an organism may affect the complexity of its regulatory networks and hence influences the specificity and the number of TFs needed for genetic regulation [20]. The number of TFs required follows a power law with the number of genes in a genome [34].

Another factor affecting the size of a predicted transcriptional network is the evolutionary distance between an organism and *B. subtilis* 168. Organisms that are taxonomically closer will contain more orthologous TFs than those which are more evolutionarily distant. Therefore, more of the transcriptional network will be predictable. For example, although the genome sizes of *B. amyloliquefaciens* (FZB42) and *B. licheniformis* (ATCC 14580) are not the biggest, these organisms are the most closely related to *B. subtilis* 168, and thus have the most predicted TFs (75 and 72 respectively). Interestingly, the absolute number of genetic regulatory relationships predicted for *B. amyloliquefaciens* and *B. licheniformis* is higher than has been observed in *B. subtilis* 168. The information about TFBSs and promoters for *B. subtilis* 168 used in BacillusRegNet is derived from DBTBS and may not represent the entire set of gene regulatory networks. Therefore, the number of binding sequences predicted for non-model organisms may be greater than those known of in *B. subtilis* 168 in BacillusRegNet. For example, there are 317 SigA promoters available for *B. subtilis* 168, but 574 SigA promoters were predicted for *B. amyloliquefaciens* (FZB42). This increase could be

Copyright 2014 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

due to incomplete coverage of the gene regulatory networks for *B. subtilis* 168, or the presence of new regulatory relationships that do not exist in *B. subtilis* 168. In addition, *B. tusciae* (DSM 2912), the most distant organism from *B. subtilis* 168 in the taxonomy, has the lowest number of predicted TFs. G*eobacillus* species have many proteins homologous to those in *B. subtilis* despite their relatively small genome sizes. Although the two *Geobacillus* species are not classified directly under the *Bacillus* taxonomy, 37 and 38 TFs respectively were predicted.

Even for model organisms, many details about genes and molecular interactions are still unknown. High-throughput experiments promise to advance the genome-scale understanding of organisms. However, the generation of wet-lab based high-throughput data is expensive and time consuming to carry out for every new species [20]. Using the RegNet system, the regulatory networks of 15 *Bacillus* species were predicted in a time- and cost-effective manner.

BacillusRegNet increases the amount of information available about host organisms for synthetic biologists. For example, transcription factors from *B. subtilis* that are not found in close relatives may be used to facilitate the engineering of regulatory pathways in the chosen non-model organism. This system also provides a valuable approach for the computational transfer of information of regulatory networks from model to non-model organisms for systems biology modelling approaches. BacillusRegNet will provide a useful resource to the biological community.

Acknowledgements

This work is supported by the Engineering and Physical Sciences Research Council under grant EP/J02175X/1 and the Engineering and Physical Sciences Research Council/National Science Foundation under grant EP/H019162/1. JB is grateful for support from the SDU2020 initiative as well as for funding from the Villum foundation. JB and RR were both supported by the Cluster of Excellence for Multimodal Computing and Interaction (MMCI) from the German Research Foundation. RR received further funding from the International Max Planck Research School in Computer Science (IMPRS). The authors would like to thank Dr. Simon Cockell, Dr. Matthew Bashton, Dr. Keith Flanagan and Prabhav Kalaghatgi for their support to setup the BacillusRegNet server.

References

[1] I. Pagani, K. Liolios, J. Jansson, I. M. Chen, T. Smirnova, B. Nosrat, V. M. Markowitz, and N. C. Kyrpides. The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Research,* 40:D571-579, 2012.

[2] V. Barbe, S. Cruveiller, F. Kunst, P. Lenoble, G. Meurice, A. Sekowska, D. Vallenet, T. Wang, I. Moszer, C. Medigue, *et al.* From a consortium sequence to a unified sequence: the *Bacillus subtilis* 168 reference genome a decade later. *Microbiology,* 155:1758-1775, 2009.

[3] C. R. Harwood. *Bacillus subtilis* and its relatives: molecular biological and industrial workhorses. *Trends in Biotechnology,* 10:247-256, 1992.

[4] L. A. Flórez, S. F. Roppel, A. G. Schmeisky, C. R. Lammers, and J. Stülke. A community-curated consensual annotation that is continuously updated: the *Bacillus subtilis* centred wiki SubtiWiki. *Database,* 2009:bap012, 2009.

Copyright 2014 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

[5] L. Westers, H. Westers, and W. J. Quax. *Bacillus subtilis* as cell factory for pharmaceutical proteins: a biotechnological approach to optimize the host organism. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research,* 1694:299-310, 2004.

[6] V. Tosato and C. V. Bruschi. Knowledge of the *Bacillus subtilis* genome: impacts on fundamental science and biotechnology. *Applied Microbiology and Biotechnology,* 64:1-6, 2004.

[7] X. H. Chen, A. Koumoutsi, R. Scholz, A. Eisenreich, K. Schneider, I. Heinemeyer, B. Morgenstern, B. Voss, W. R. Hess, O. Reva, *et al.* Comparative analysis of the complete genome sequence of the plant growth-promoting bacterium *Bacillus amyloliquefaciens* FZB42. *Nature Biotechnology*, 25:1007-1014, 2007.

[8] M. Eppinger, B. Bunk, M. A. Johns, J. N. Edirisinghe, K. K. Kutumbaka, S. S. K. Koenig, H. Huot Creasy, M. J. Rosovitz, D. R. Riley, S. Daugherty, *et al.* Genome Sequences of the Biotechnologically Important *Bacillus megaterium* Strains QM B1551 and DSM319. *Journal of Bacteriology,* 193:4199-4213, 2011.

[9] M. Rey, P. Ramaiya, B. Nelson, S. Brody-Karpin, E. Zaretsky, M. Tang, A. de Leon, H. Xiang, V. Gusti, I. G. Clausen, *et al.* Complete genome sequence of the industrial bacterium *Bacillus licheniformis* and comparisons with closely related *Bacillus* species. *Genome Biology,* 5:r77, 2004.

[10] H. Takami, Y. Takaki, G.-J. Chee, S. Nishi, S. Shimamura, H. Suzuki, S. Matsui, and I. Uchiyama. Thermoadaptation trait revealed by the genome sequence of thermophilic *Geobacillus kaustophilus*. *Nucleic Acids Research*, 32:6292-6303, 2004.

[11] Y. J. Tang, R. Sapra, D. Joyner, T. C. Hazen, S. Myers, D. Reichmuth, H. Blanch, and J. D. Keasling. Analysis of metabolic pathways and fluxes in a newly discovered thermophilic and ethanol-tolerant *Geobacillus* strain. *Biotechnology and Bioengineering,* 102:1377-1386, 2009.

[12] J. Baumbach. On the power and limits of evolutionary conservation—unraveling bacterial gene regulatory networks. *Nucleic Acids Research,* 38:7877-7884, 2010.

[13] M. Das and H.-K. Dai. A survey of DNA motif finding algorithms. *BMC Bioinformatics,* 8:S21, 2007.

[14] M. L. Bulyk, A. M. McGuire, N. Masuda, and G. M. Church. A Motif Co-Occurrence Approach for Genome-Wide Prediction of Transcription-Factor-Binding Sites in Escherichia coli. *Genome Research*, 14:201-208, 2004.

[15] T. Venancio and L. Aravind. Reconstructing prokaryotic transcriptional regulatory networks: lessons from actinobacteria. *Journal of Biology,* 8:29, 2009.

[16] S. A. F. T. van Hijum, M. H. Medema, and O. P. Kuipers. Mechanisms and Evolution of Control Logic in Prokaryotic Transcriptional Regulation. *Microbiology and Molecular Biology Reviews,* 73:481-509, 2009.

[17] D. A. Reddy, B. V. L. S. Prasad, and C. K. Mitra. Comparative analysis of core promoter region: Information content from mono and dinucleotide substitution matrices. *Computational Biology and Chemistry,* 30:58-62, 2006.

[18] M. Beckstette, R. Homann, R. Giegerich, and S. Kurtz. Fast index based algorithms and software for matching position specific scoring matrices. *BMC Bioinformatics,* 7:389, 2006.

Copyright 2014 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

[19] T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble. MEME Suite: tools for motif discovery and searching. *Nucleic Acids Research,* 37:W202-W208, 2009.

[20] J. Baumbach, T. Wittkop, C. K. Kleindt, and A. Tauch. Integrated analysis and reconstruction of microbial transcriptional gene regulatory networks using CoryneRegNet. *Nature Protocols,* 4:992-1005, 2009.

[21] J. Oberto. FITBAR: a web tool for the robust prediction of prokaryotic regulons. *BMC Bioinformatics,* 11:554, 2010.

[22] P. S. Novichkov, D. A. Rodionov, E. D. Stavrovskaya, E. S. Novichkova, A. E. Kazakov, M. S. Gelfand, A. P. Arkin, A. A. Mironov, and I. Dubchak. RegPredict: an integrated system for regulon inference in prokaryotes by comparative genomics approach. *Nucleic Acids Research,* 38:W299-W307, 2010.

[23] J. Baumbach, T. Wittkop, K. Rademacher, S. Rahmann, K. Brinkrolf, and A. Tauch. CoryneRegNet 3.0-An interactive systems biology platform for the analysis of gene regulatory networks in corynebacteria and *Escherichia coli*. *Journal of Biotechnology,* 129:279 - 289, 2007.

[24] J. Baumbach. CoryneRegNet 4.0 - A reference database for corynebacterial gene regulatory networks. *BMC Bioinformatics,* 8:429, 2007.

[25] J. Baumbach, S. Rahmann, and A. Tauch. Reliable transfer of transcriptional gene regulatory networks between taxonomically related organisms. *BMC Systems Biology,* 3:8, 2009.

[26] J. Baumbach, K. Brinkrolf, L. Czaja, S. Rahmann, and A. Tauch. CoryneRegNet: An ontology-based data warehouse of corynebacterial transcription factors and regulatory networks. *BMC Genomics,* 7:24, 2006.

[27] J. Krawczyk, T. A. Kohl, A. Goesmann, J. Kalinowski, and J. Baumbach. From *Corynebacterium glutamicum* to *Mycobacterium tuberculosis*—towards transfers of gene regulatory networks and integrated data analyses with MycoRegNet. *Nucleic Acids Research,* 37:e97, 2009.

[28] N. Sierro, Y. Makita, M. de Hoon, and K. Nakai. DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. *Nucleic Acids Research,* 36:D93-D96, 2008.

[29] G. Misirli. Data integration strategies for informing computational design in synthetic biology. Ph.D., School of Computing Science, Newcastle University, UK, 2013.

[30] G. Misirli, A. Wipat, J. Mullen, K. James, M. Pocock, W. Smith, N. Allenby, and J. Hallinan. BacillOndex: An Integrated Data Resource for Systems and Synthetic Biology. *Journal of Integrative Bioinformatics,* 10:224, 2013.

[31] J. Kohler, J. Baumbach, J. Taubert, M. Specht, A. Skusa, A. Ruegg, C. Rawlings, P. Verrier, and S. Philippi. Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics,* 22:1383-1390, 2006.

[32] P. S. Dehal, M. P. Joachimiak, M. N. Price, J. T. Bates, J. K. Baumohl, D. Chivian, G. D. Friedland, K. H. Huang, K. Keller, P. S. Novichkov, *et al.* MicrobesOnline: an integrated portal for comparative and functional genomics. *Nucleic Acids Research,* 38:D396-D400, 2009.

Copyright 2014 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).

[33] J. D. Helmann and J. C. P. Moran. RNA Polymerase and Sigma Factors. in *Bacillus subtilis and its closest relatives: from genes to cells. ASM Press, Washington, DC*, ed: Amer Society for Microbiology, 2002, pp. 289-312.

[34] V. Charoensawan, D. Wilson, and S. A. Teichmann. Genomic repertoires of DNA-binding transcription factors across the tree of life. *Nucleic Acids Research*, 38:7364-7377, 2010.

Copyright 2014 The Author(s). Published by Journal of Integrative Bioinformatics.
This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License (http://creativecommons.org/licenses/by-nc-nd/3.0/).