Contents lists available at ScienceDirect

BBA - Gene Regulatory Mechanisms

journal homepage: www.elsevier.com/locate/bbagrm

AURTHO: Autoregulation of transcription factors as facilitator of *cis*-acting element discovery

Sinaeda Anderssen[a], Aymeric Naômé[a, b], Cédric Jadot[a], Alain Brans[a], Pierre Tocquin[b, c], Sébastien Rigali[a, b, *]

[a] *InBioS – Center for Protein Engineering, University of Liège, B-4000 Liège, Belgium*
[b] *HEDERA 22, Boulevard du Rectorat 27b, B-4000 Liège, Belgium*
[c] *InBioS – PhytoSystems, University of Liège, B-4000 Liège, Belgium*

ARTICLE INFO

Keywords:
Autoregulation
Transcription factor binding site discovery
Regulon prediction
cis-Acting elements
Transcription control

ABSTRACT

Transcriptional regulation is key in bacteria for providing an adequate response in time and space to changing environmental conditions. However, despite decades of research, the binding sites and therefore the target genes and the function of most transcription factors (TFs) remain unknown. Filling this gap in knowledge through conventional methods represents a colossal task which we demonstrate here can be significantly facilitated by a widespread feature in transcriptional control: the autoregulation of TFs implying that the yet unknown transcription factor binding site (TFBS) is neighboring the TF itself. In this work, we describe the "AURTHO" methodology (AUtoregulation of oRTHOlogous transcription factors), consisting of analyzing upstream regions of orthologous TFs in order to uncover their associated TFBSs. AURTHO enabled the *de novo* identification of novel TFBSs with an unprecedented improvement in terms of quantity and reliability. DNA-protein interaction studies on a selection of candidate *cis*-acting elements yielded an >90 % success rate, demonstrating the efficacy of AURTHO at highlighting true TF-TFBS couples and confirming the identification in a near future of a plethora of TFBSs across all bacterial species.

1. Introduction

In the prokaryotic world, subtle changes in the environment can have limited or instead widespread effects on the expression of genes, permitting an efficient response to new conditions. This modulation of gene expression is mediated by different mechanisms, the best-known of which involves transcription factors (TF) that, through binding of specific DNA sequences will activate or inhibit the transcription of target genes. Regulators often control the expression of multiple genes by binding to similar transcription factor binding sites (TFBS) upstream of each of its targeted genes or transcription units [1–4]. Despite having been in the spotlight the longest among all regulation mechanisms, a great deal of mystery still pertains to transcriptional networks even in well-studied microorganisms like *Escherichia coli* [5,6]. Actually, in most bacteria, only a handful of TFs have been studied, revealing just an inkling of the regulatory networks they use to control cellular processes and adapt to their environment rapidly and efficiently.

Using a wet lab approach, unveiling novel TF-TFBS couples and their regulatory network can take years, but high throughput approaches such as RNA-seq, ChIP-Seq, and DAP-seq have been game changers in regulation data acquisition [5,7–11]. These approaches largely facilitate the assessment of the transcriptional output in response to a specific set of signals. However, researchers are often limited and biased when testing a set of laboratory culture conditions, which rarely reflect the bacteria's natural environment. Indeed, the transcriptional response, and therefore the binding of TFs, is also a dynamic process that highly varies in time and space according to the state of growth or the step of the life cycle for bacteria that undergo extensive physiological and morphological differentiations [12]. Therefore, the fraction of TFs which are only expressed and needed in very specific conditions are unlikely to be highlighted *via* these studies.

Completely different approaches starting from *in silico* analyses have also been used. Usually, these approaches first acquire the knowledge of the TFBS, from which the regulon can be inferred, after which its function can be deduced through the analysis of the target genes' func-

Abbreviations: TF, transcription factor; TFBS, transcription factor binding site; EMSA, electromobility gel shift assay

[*] Corresponding author at: InBioS – Center for Protein Engineering, University of Liège, B-4000 Liège, Belgium.
E-mail address: srigali@uliege.be (S. Rigali).

https://doi.org/10.1016/j.bbagrm.2022.194847
Received 2 May 2022; Received in revised form 4 July 2022; Accepted 18 July 2022
1874-9399/© 20XX

tions [4,13–17]. With the advent of genome sequencing technologies, researchers have been working to exploit these data to uncover conserved regulatory elements and link them to a TF. As early as 2002, the genomes of three model micro-organisms; *E. coli*, *Bacillus subtilis*, and *Streptomyces coelicolor*, had been studied with the aim to uncover overrepresented dyad-type motifs in intergenic regions of the genome, where *cis*-acting elements are expected to be found [18–20]. During that same period, we used an *in silico*-based approach to show that refining the classification of TFs into sub-families beyond the sequence of their helix-turn-helix motif facilitates the discovery of their binding sites. In addition, this work demonstrated that using the autoregulatory property of bacterial regulators in an *in silico* approach was an effective way to assign a discovered TFBS to its cognate TF [15,21]. Now that the number of available genomes has significantly grown, approaches based on comparative genomics and more specifically on phylogenetic footprinting, have become possible [16,22,23]. Phylogenetic footprinting is a method that aims at discovering conserved regulatory sequences in orthologous UTRs (UnTranslated Region) in different genomes, as it is believed that functional features are encoded in evolutionarily conserved DNA sequences. Thus, the traits that are targeted are regulatory DNA sequences (TFBSs) and their associated TF. The research group of Prof. Rodionov has indeed shown through their “regulon propagation and reconstruction” approach that certain orthologous TFs and their cognate TFBSs are conserved across an extensive variety of taxa [16,24–28].

We predict that this type of approach, when used on a more closely related taxonomic group, will prove to be even more prolific in terms of the quantity of discovered *cis-trans* relationships. Indeed, numerous TF-TFBS couples are only conserved between closely related species, and this focused approach will likely point out taxon-specific regulatory interactions. With this in mind, we developed a *de novo* approach and assessed the extent to which it could accelerate the discovery of DNA sequences recognized by TFs. In contrast to previously used comparative genomics *in silico* approaches, our methodology draws on a widespread property of TFs, *i.e.*, they often control their own expression, which imposes that the location of the searched TFBS is in the close vicinity of the TF gene itself. Combined with the conservation of the TFBS between orthologous TFs, this guided the development of the AURTHO methodology (AUtoregulation of oRTHOlogous transcription factors), consisting of analyzing upstream regions of orthologous TFs in order to uncover their associated TFBSs.

As a case study to test the AURTHO methodology, we focused our attention on one family of TFs, the LacI family, and selected a closely related taxon, the *Streptomyces* genus, as the latter has been shown to encode large numbers of TFs (12.3 % of the model species' genome is dedicated to encoding regulatory genes) [29]. The AURTHO strategy revealed to be extremely efficient at providing reliable candidate TFBSs as the presented work not only confirmed the TFBS of the five LacI-TFs already studied in streptomycetes but also proposed a cognate TFBS for 90 additional and yet uncharacterized LacI-TFs thereby largely filling the gap in knowledge about *cis*-acting elements. As autoregulation is a feature of many different TF families, our results suggest that the application of the AURTHO approach across all bacterial species will highly facilitate the discovery of novel TF-TFBS couples.

2. Results and discussion

2.1. Starting hypotheses and the AURTHO methodology

The *de novo* approach used to unveil the TFBSs of LacI-family TFs is based on three main assumptions: (i) orthologous TFs bind to identical motifs on DNA, (ii) LacI TFs often (70 % according to Ravcheev et al., 2014) regulate their own expression (autoregulation), meaning their binding site can be found in the upstream region of the gene encoding them, and (iii) its primary target gene(s) is (are) usually found adjacent to or in the same transcriptional unit as that of the TF, reinforcing the probability of finding its binding site in close vicinity to the TF gene. Additionally, for members of the LacI-family of TFs (used in this work as a case study) the binding sites are easily spotted as they are usually characterized by palindromic sequences of even length which contain a typical CG-pair at the centre of the motif [28]. Nonetheless, some atypical binding sites have been identified, showcasing uneven lengths, the absence of a CG-pair in the centre [30], directed repeats [31] and/or a stretch of less conserved nucleotides of variable length between the two inverted repeats (though for a single TF and its orthologs, the length is usually conserved) [28].

The methodology that guided our approach is detailed in the flowchart presented in Fig. 1. First, genomes from the genus *Streptomyces* were downloaded from the NCBI database and filtered to retrieve only the ones annotated as “Complete” in their assembly status (assembly.info on GitHub). Proteinortho [32] was used to create clusters of orthologous genes (COGs) by performing diamond blast in an all-*versus*-all manner, and clustering genes using a reciprocal best alignment heuristic (RBAH). Simultaneously, an hmmscan using HMMER3 was performed on all genomes against the Pfam-A profile database [33–35], and TF genes were classified into families through signature domain combinations, as described in the P2TF database [36]. For the LacI-family of TFs, the signature domain combination consists of a LacI DNA-binding domain (PF0356) and a periplasmic binding protein domain (PF0532, PF13377 or PF13407) [28]. However, according to the P2TF database, the presence of a LacI-HTH motif inside the DNA-binding domain is a sufficient predictor of a protein belonging to this family of TFs [36]. For every gene identified as a LacI TF, we extracted the COG they belonged to, and “manually” checked for functional coherence inside the COG based on the gene annotations. Only COGs in which most annotations were coherent with a regulatory function were conserved for further analysis. For each of them, the upstream sequences of the LacI-TF genes were extracted, the length of which is variable as the extraction halted as soon as the translational start/stop codon of an upstream gene was encountered. Different maximum lengths of search regions were tested (500 bp, 300 bp, 100 bp with an additional 50 bp inside the coding region). For each LacI-COG, these sequences were aligned with the MEME software [37] using two different search parameters termed ZOOPS (Zero or One Occurrence Per Sequence) and ANR (Any Number of Repetitions), and three different search lengths (small = 10 nucleotides (nt), medium = 20 nt, and long = 30 nt). MEME produced four motifs per search, and the results for each combination of parameters were manually curated to identify sites that were most consistent with characteristics of known LacI binding sites, namely the palindromic property of the site and the central CG-pair [28]. Finally, the FASTA-format matrices of putative binding sites were used to create sequence logos with WebLogo [38] and to design Cy5-marked DNA probes containing the consensus binding site for each LacI-COG. A series of LacI-family TFs were selected to experimentally validate the predicted DNA-protein interaction through Electrophoretic Mobility Shift Assays (EMSAs). Finally, an additional round of manual inspection was performed in COGs' cases which required manual inspection of the gene locus organization in order to extract the proper gene's upstream region (see step 8 in Figs. 1 and 2).

2.2. De novo identification of binding sites of LacI-family TFs in streptomycetes

2.2.1. LacI-family transcription factor identification

LacI-family TFs were identified by the presence of a typical LacI helix-turn-helix motif (PF0356) in the N-terminal DNA-binding domain of the protein sequence. As expected for the *Streptomyces* genus, in which sugar catabolism regulation is essential for adaptation to diverse environments [39,40], LacI TFs were identified in all 182 studied complete genomes (Supplementary Fig. S1). However, there was a great disparity

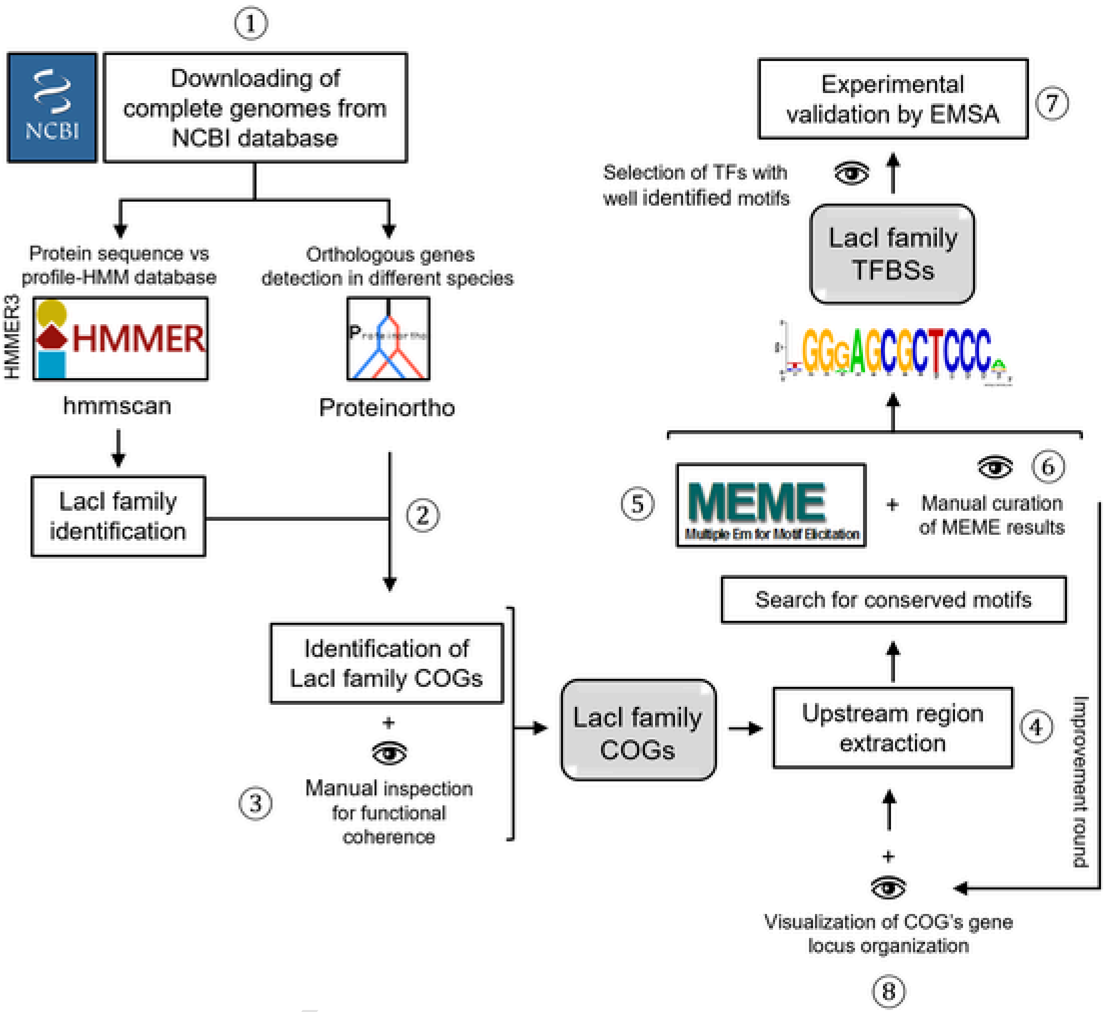

1
NCBI
Downloading of complete genomes from NCBI database
Protein sequence vs profile-HMM database
Orthologous genes detection in different species
HMMER3
HMMER
hmmscan
Proteinortho
LacI family identification
2
Identification of LacI family COGs
+
Manual inspection for functional coherence
3
LacI family COGs
Upstream region extraction
4
+
Visualization of COG's gene locus organization
8
Improvement round
Search for conserved motifs
5
MEME
Multiple Em for Motif Elicitation
+
Manual curation of MEME results
6
GGGAGCGCTCCC
LacI family TFBSs
Selection of TFs with well identified motifs
Experimental validation by EMSA
7

Fig. 1. Flowchart illustrating the different steps of the AURTHO approach. The "eye" icon indicates the different steps that required manual inspection of the software/algorithm output before proceeding to the following step(s) of the methodology and/or to improve/increase the quality/quantity of the data generated. Step 1. Downloading complete genomes of interest from de NCBI database, identifying the LacI regulators through hmmscan, and clustering genes into orthologous groups; Step 2. Selecting the COGs that contain at least one LacI TF (identified by hmmscan); Step 3. Verification that the functional annotation of the genes in these COGs are coherent with a regulatory role; Step 4. For LacI-COGs, extraction of the upstream region of TF genes; Step 5. Alignment using the MEME software of the sets of upstream regions for each COG; Step 6. Manual curation of MEME results to identify over-represented motifs also containing typical characteristics of LacI-family TFBSs; Step 7. Selection of predicted motifs for experimental validation through EMSAs; Step 8. Improvement round for COGs with no proposed motif, involving the visualization of the gene locus organization and extraction of the region upstream of the first gene of the transcription unit that contain the TF of interest.

in the number of LacI regulators identified. *Streptomyces bingchenggensis* (BCW-1) possesses 69 LacI TFs, while *Streptomyces olivoreticuli* (subsp. *olivoreticuli* strain = ATCC 31159) only encodes 6 LacI genes (Fig. S1). This goes far beyond any explanation related to their genomes' size, as there is no correlation between the length of the chromosome and the relative abundance of LacI TFs (11.9 Mb/9692 genes and 8.8 Mb/7102 genes for *S. bingchenggensis* and *S. olivoreticuli*, respectively).

In total, in 182 *Streptomyces* strains, 4403 LacI TFs were identified, grouped into 167 COGs. Among these, only 5 (~3 % of all LacI TFs) have been subject to studies in *Streptomyces* species, *i.e.*, i) the galactomannan/mannobiose/mannose utilization repressor ManR (LacI003 in Table 1, conserved in 177/182 species) [41], (ii) the maltose/maltodextrin catabolism pathway regulator MalR (LacI005 in Table 1, conserved in 176/182 species) [31,42–45], iii) the cellulose/cello-oligosaccharide utilization regulator CebR (LacI006 in Table 1, conserved in 153/182 species) [46–50], iv) the xylan/xylo-oligosaccharide utilization repressor BxlR (LacI015 in Table 1, conserved in 88/182 species) [30,51,52], and v) the agar-utilization regulator DagR (LacI139 in Table 1) [53], the latter being one of the rarest LacI TF, only conserved in two *Streptomyces* species. Strikingly, the function of the two most conserved LacI TFs (LacI001 and LacI002 in Table 1) is unknown, further illustrating the lack of knowledge about transcriptional regulation in this well-studied bacterial genus. Remarkably, 25 LacI TFs were only present in one single species, meaning they were part of "orphan" COGs containing only that single gene. In these cases, it is inherently impossible to perform a comparative genomics approach, which requires the comparison of two or more sequences.

2.2.2. *Identification of TF binding sites*

For each the 167 LacI-family COGs, a set of upstream regions was extracted with varying lengths as described in the Methodology section. This resulted in 138 sets of two or more upstream regions. Indeed, in the remaining cases, the COG was either orphan (one gene), or there was either only one, or no gene in the COG for which an upstream region was present. This happens when the TF is co-transcribed with other genes in its transcription unit. As explained above, three maximum lengths of upstream sequences were tested for the MEME analysis, but overall, a maximum length of 300 bp (halted whenever an upstream coding region was encountered) yielded the best results in terms of number of discovered motifs and their resolution (Supplementary

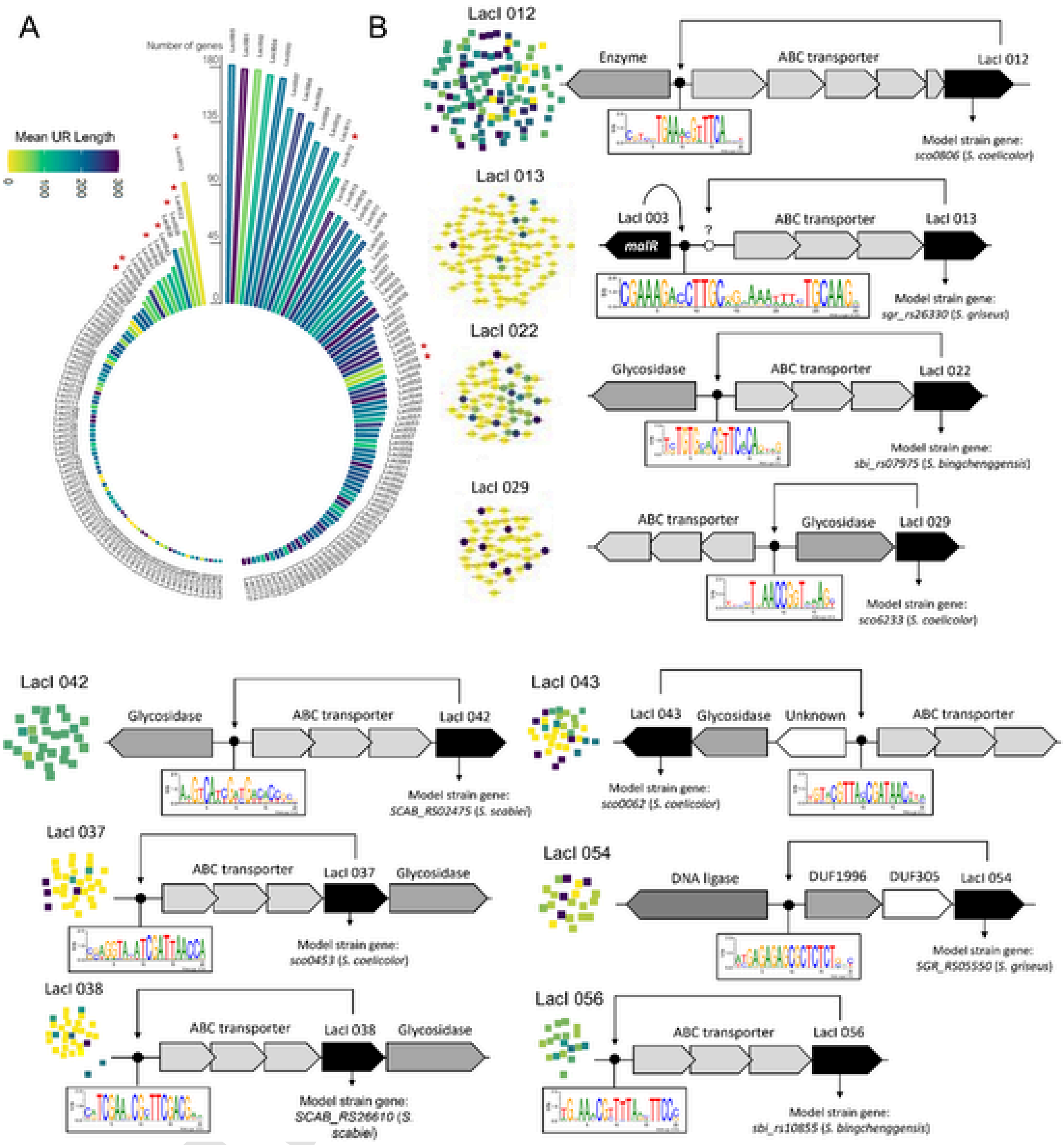


Fig. 2. Inspection of the length of the TF upstream region and the gene locus organization. (A) Average length of upstream regions for each COG. The height of the bars is relative to the number of members in each COG, and the color indicates the mean length of extracted upstream regions. The right part of the barplot shows the TF COGs for which a cognate TFBS was predicted, while the left part corresponds to those that yielded no potential motif. Based on these data, we selected COGs (red asterisks) for which the average length of UPS region was low and analyzed the operon organization for each member of the COG. (B) Operon organization of COGs that displayed a low average upstream region length, for which we did or not find a putative binding site. The node cluster next to the operon organization represents the corresponding COG, where each node is a gene of the COG, and the color indicates the length of the upstream region of that particular gene. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

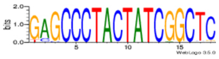
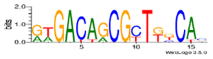
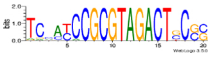
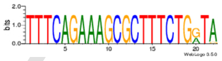
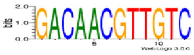
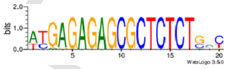
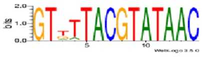
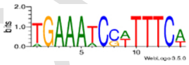
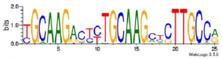
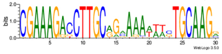
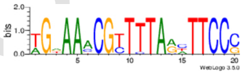
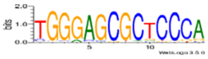
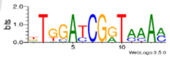
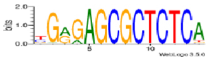
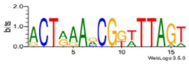
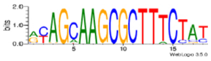
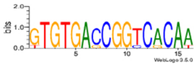
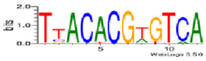
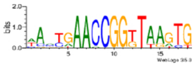
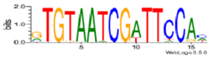
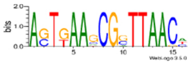
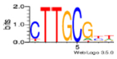
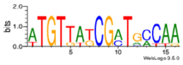
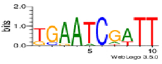
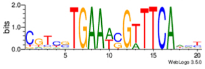
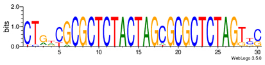
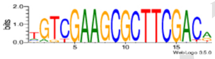
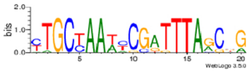
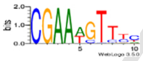
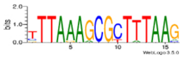
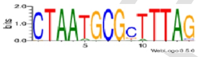
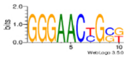
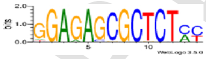
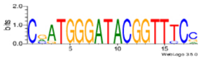
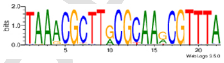
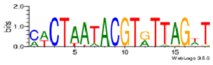
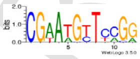
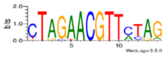
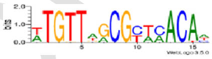
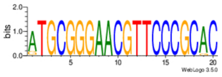
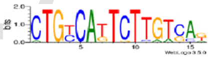
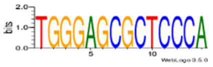
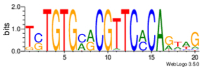
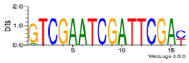
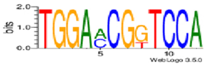
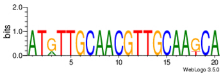
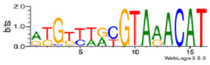
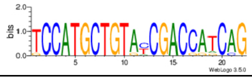
Fig. S2). This was supported by the previous observation of Ravcheev et al. [28] that LacI binding sites are rarely found beyond 300 nucleotides upstream of the target gene, or after the beginning of the coding region.

In order to first assess the reliability of our *de novo* approach, we singled out the studied LacI regulators (ManR, MalR, CebR, BxlR, and DagR), and checked if the motifs we generated using our *in-silico* approach correspond to their experimentally determined *cis*-acting sequences. As presented in Table 1, for ManR (LacI003), CebR (LacI006), and BxlR (LacI015), the *de novo* identified motifs were identical to their experimentally identified consensus sequences, *i.e.*, GACAACGTTGTC [41], TGGGAGCGCTCCCA [50], and CGAA-Nx-TTCG [30,51,52], respectively. For MalR (LacI005), the two binding sites deduced by DNase footprinting assays [31] were also found (see Table 1), further confirming that our approach is appropriate for deducing over-represented motifs that closely relate to the ones that were experimentally identified.

In the case of DagR, its DNA-binding site was not identified during our first manual inspection of MEME-generated motifs. Indeed, this TF is only present in two strains (*S. coelicolor* and *S. bingchenggensis*), meaning there were only two upstream regions to align. In this case, MEME is often not able to distinguish motifs found by chance from potentially biologically significant ones, causing proposed motifs to have very high E-values. Hence, it was only upon re-examination of the four motifs proposed by MEME that we identified the one that corresponded to one of the validated binding sites of DagR (LacI139), AACCGGTT [53].

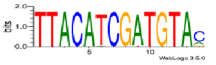
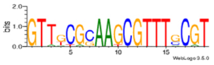
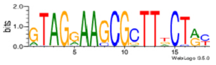
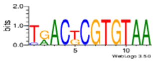
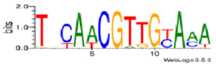
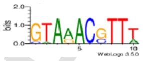
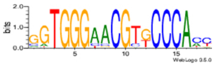
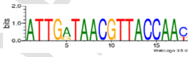
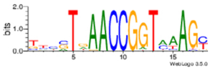
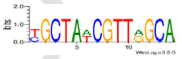
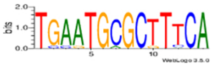
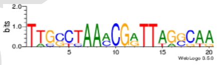
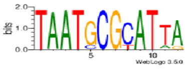
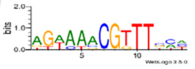
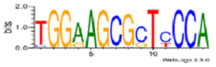
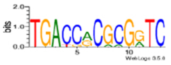
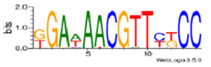
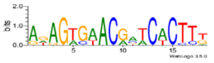
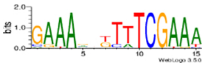
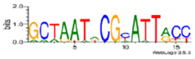
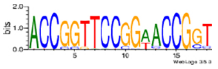
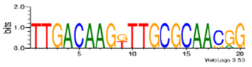
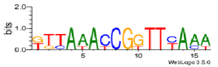
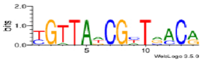
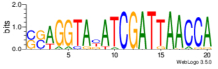
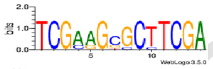
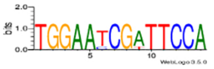
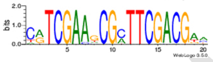
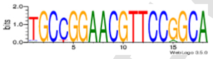
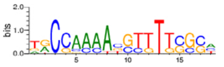
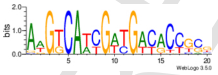
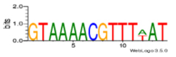
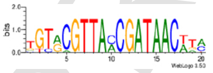
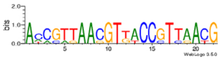
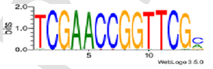
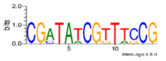
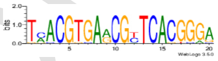
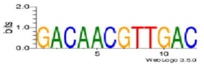
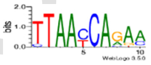
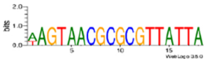
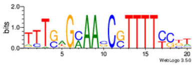
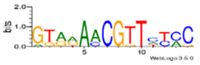
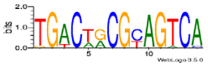
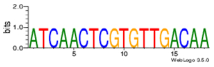
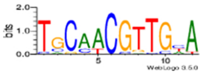
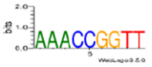
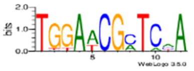
Of the 133 unstudied LacI-COGs for which two or more upstream sequences could be extracted, one or two putative binding site(s) in their upstream region was found for 82 (~62 %) of them (Table 1). In addition, 9 motifs were further identified (6) or improved (3) by extracting the upstream region of the first gene of a transcriptional unit (operon) that contains the TF gene (see below in the next section), bringing the

Table 1
LacI COGs and their AURTHO predicted binding sites.

LacI COG (repr. memb.)	Predicted TFBS (WebLogo)	Occur. (%)	LacI COG (repr. memb.)	Predicted TFBS (WebLogo)	Occur. (%)
001 (B) (SCO3943)		181 (99.5)	052 (A) (T261_RS40815)		21 (11.5)
002 (B) (SCO4158)		181 (99.5)	053 (A) (SGR_RS04505)		20 (11)
003, *manR* (SCO1078)		177 (97.3)	054 (A*) (SGR_RS05550)		20 (11)
004 (A) (SCO1642)		176 (96.7)	055 (B) (NI25_RS02935)		17 (9.3)
005, *malR* (SCO2232)	 	176 (96.7)	056 (B*) (SBI_RS10855)		17 (9.3)
006, *cebR* (SCO2794)		153 (84.1)	057 (B) (SBI_RS36280)		17 (9.3)
007 (A) (SCO6713)		152 (83.5)	058 (A) (SBI_RS03995)		16 (8.8)
008 (A) (SCO2753)		148 (81.3)	059 (A) (SBI_RS36130)		16 (8.8)
009 (A) (SCO0886)		135 (74.2)	060 (A) (SFLA_RS18915)		16 (8.8)
010 (A) (SCO2745)		133 (73.1)	061 (A) (T261_RS24440)		16 (8.8)
011§ (C) (SCO7014)		112 (61.5)	062 (B) (SCO6349)		15 (8.2)
012 (A, A*) (SCO0806)	 	101 (55.5)	064 (B) (SBI_RS03795)		14 (7.7)
014 (A) (SCO5692)		88 (48.4)	065 (A) (XNR_RS03065)		14 (7.7)
015, *bxlR1* (SCO7027)		88 (48.4)	066 (A) (SBI_RS46800)		13 (7.1)
016 (A) (SCO0953)		87 (47.8)	069 (C) (SBI_RS10490)		12 (6.6)
017 (A) (SCO6598)		83 (45.6)	070 (C) (SBI_RS48025)		12 (6.6)
018 (A) (SCO1956)		80 (44)	071 (A) (SCAB_RS41390)		12 (6.6)
019 (C) (SCO1376)		76 (41.8)	072 (A) (STRVI_RS24580)		12 (6.6)
020 (B) (SBI_RS40970)		62 (34.1)	073 (A) (SVTN_RS35145)		12 (6.6)
021 (C) (SCO6986)		60 (33)	074 (A) (SVTN_RS32805)		11 (6)
022 (A*) (SBI_RS07975)		58 (31.9)	076 (A) (SBI_RS02810)		10 (5.5)
023 (A) (SBI_RS45370)		57 (31.3)	077 (A) (SBI_RS06345)		10 (5.5)
024 (A) (SCO7411)		50 (27.5)	078 (C) (SBI_RS08340)		10 (5.5)

(*continued on next page*)

Table 1 (*continued*)

LacI COG (repr. memb.)	Predicted TFBS (WebLogo)	Occur. (%)	LacI COG (repr. memb.)	Predicted TFBS (WebLogo)	Occur. (%)
025 (A) (SCO7502)		50 (27.5)	079 (A) (SBI_RS42795)		10 (5.5)
026 (A) (SBI_RS08050)		48 (26.4)	080 (A) (AVL59_RS26565)		9 (4.9)
027 (A) (SCO1066)		48 (26.4)	083 (A) (SFLA_RS00305)		8 (4.4)
028 (A) (SCO7554)		48 (26.4)	084 (A) (SHJGH_RS07625)		8 (4.4)
029 (A*) (SCO6233)		45 (24.7)	085 (A) (A4E84_RS39220)		7 (3.8)
031 (A) (SGR_RS11925)		38 (20.9)	086 (A) (AA958_RS04325)		7 (3.8)
032 (A) (SCO0629)		36 (19.8)	090 (A) (SCAB_RS08895)		7 (3.8)
033 (A) SBI_RS48885		34 (18.7)	091 (B) (SCAB_RS37085)		7 (3.8)
034 (A) (SCAB_RS41450)		33 (18.1)	093 (A) (SXIM_RS01320)		7 (3.8)
035 (C) (SCO0289)		32 (17.6)	094 (A) (SXIM_RS22465)		7 (3.8)
036 (A) (SBI_RS46210)		31 (17)	096 (B) (SBI_RS08910)		6 (3.3)
037 (A, B*) (SCO0456)		31 (17)	097 (A) (SBI_RS31600)		6 (3.3)
					
038 (A, A*) (SCAB_RS26610)		30 (16.5)	101 (A) (STRVI_RS14955)		6 (3.3)
					
039 (A) (SGR_RS17280)		30 (16.5)	102 (A) (SVTN_RS01870)		6 (3.3)
042 (A*) (SCAB_RS02460)		26 (14.3)	103 (A) (SVTN_RS03670)		6 (3.3)
043 (A*) (SCO0062)		26 (14.3)	106 (B) (SCAB_RS42505)		5 (2.7)
044 (A) (SBI_RS01965)		25 (13.7)	107 (B) (SCO0360)		5 (2.7)
046 (A) (SBI_RS48670)		24 (13.2)	110 (A) (AS200_RS41850)		4 (2.2)
047 (C) (STRVI_RS12305)		24 (13.2)	112 (A) (CFP59_RS47970)		4 (2.2)
048 (C) SBI_RS03890		23 (12.6)	114 (A) (SBI_RS10425)		4 (2.2)
049 (A) (SCAB_RS03420)		23 (12.6)	117 (A) (SCAB_RS06320)		4 (2.2)
050 (A) (WQO_RS32820)		23 (12.6)	139, *dagR*† (SCO3485)		2 (1.1)
051 (B) (SBI_RS21665)		22 (12.1)			

◀ The COGs are numbered from 1 to 142, by decreasing number of strains represented in the COG. Orphan do not figure in this table. The locus tag under the COG number is that of the representative gene (the gene belonging to the most studied species possessing a gene belonging to this COG). The letter associated with each COG indicates the "reliability group" for the predicted motif, *i.e.* A) highly reliable (palindrome sequence with a central CG pair), B) reliable but with either incomplete palindromic sequence or missing the central CG pair, and C) atypical LacI sites (CG pair not conserved and no inverted repeat). The occurrence represents the number of different strains that are represented in each COG (182 strains were studied). Symbols: §, indicates that the motifs have been found more than once and could thus be part of a direct repeat; *, indicates that the motif was identified by the manual inspection of the TF gene locus organization (step 8 in the flowchart of Fig. 1). †, indicates the motif did not pass the threshold but was found by manual inspection based on the known binding site retrieved from the literature survey (work on DagR [53]).

total number of COGs with a predicted TFBS to 88 (~66 %). Based on the previously defined characteristics of LacI TFBSs (central CG pair and inverted repeat sequence), we defined different "reliability groups" for the predicted motifs (categories A, B, and C in Table 1) we think reflect the probability of the site being bound by its cognate TF. For example, the TGTGACCGGTCACA conserved motif found upstream of LacI059 orthologs presents of 14 bp perfect inverted repeat centred on a CG pair. For over 70 % of LacI-COGs, the predicted motif is considered to be highly reliable (assigned A in Table 1), as they possess both characteristics. TF-TFBS couples have a lower predicted reliability if one of these two characteristics is missing, which was the case for 11 LacI-COGs (assigned B in Table 1). This is for instance the case of the predicted motif of LacI 001 and LacI 002, the first of which, although containing an inverted repeat (GAGCC-N8-GGCTC), lacks the typical central CG-pair, and the second on the other hand possessing the central CG pair but for which the left part of the motif does not at all reflect any kind of symmetry with the right part. For the remaining 9 LacI-COGs, the best motif does not possess either of these two sequence features and, consequently, they have a much lower confidence score (motifs assigned C in Table 1).

2.3. Improvement round by inspection of TF genetic locus organization

Around 40 % of LacI-COGs did not yield any potential binding site using our approach (see Fig. 2A). In most cases (LacI120–LacI142 in Fig. 2A) the size of the COG was so small (2 or 3 members in the COG) that, as demonstrated with the DagR example discussed above, MEME likely could not distinguish motifs occurring by chance from biologically significant ones. Indeed, usually, when the number of representatives of one COG is too small, the entire region upstream of the TF is conserved which prevents the identification of the functional conserved *cis*-acting elements. Nonetheless, there is a number of COGs for which we unexpectedly did not find an over-represented motif. Although this could simply be due to the lack of autoregulation for these COGs, further investigation revealed that in some cases, the average length of the region upstream of these COGs was smaller than for COGs for which we could find a conserved motif (Fig. 2A). Indeed, LacI-TFs are typically encoded in the divergent direction of the genes of the operon they regulate, and through binding to the *cis*-acting element in the intergenic region between its own gene and the upstream gene, it can control both transcription units in concert. However, the genetic organization is not always as such, and the TF can sometimes be found in between other genes belonging to the same transcription unit or even in the last position of the latter. Fig. 2B illustrates the LacI COGs where the operon organization clearly prevented the identification of a binding site in the upstream region of the TF encoding gene. In these cases, the TF is still likely to bind to the region upstream of the sets of genes that constitute the whole transcription unit to which the TF encoding gene belongs to. Therefore, the correct search region is not in the upstream region of the TF gene, but in the transcription unit's upstream region.

With this in mind, we selected the COG that is mostly present in the first position of the transcription unit in order to repeat the upstream region extraction and the MEME analysis. The selected examples where this additional round allowed the identification of a conserved motif or to modify the motif originally found are presented in Fig. 2B. Notably, for six of the selected examples, this additional round of manual inspection allowed to identify 5 class A motifs (022, 029, 042, 043, and 054) and one class B motif (056) (Fig. 2B and Table 1). The remaining three examples involve COGs for which a motif was discovered through the direct extraction of the TF gene upstream region (012, 037 and 038). However, this additional round enabled the improvement of two of the motifs (for 012 and 038), and the identification of a second, binding site for LacI 037 which, although it contains a well-conserved CG-pair in the centre, the left part of the palindrome can only be guessed from the sequence logo, classifying this motif in the B category. In this case, the additional round brought more ambiguity to the predicted TFBS, and which one is the true binding site for LacI 037 remains to be determined. For LacI 012 and LacI 038, the motifs that MEME proposed were very similar to the ones uncovered the first time. Hence, this further strengthens our confidence in the palindromic sequence that was initially found. Finally, among the other COGs that were selected, LacI 013 represents a very peculiar case as it is part of the *malEFG* operon divergently transcribed from the gene encoding MalR (belonging to the COG LacI 005). As a consequence, the examination of this operon's regulatory region only highlighted the MalR binding site again, with LacI 013 possibly competing for the same site or targeting a site residing elsewhere in the chromosome. Nonetheless, this additional manual check remains essential in cases where the operon's organization deviates from the "typical" topology. This enabled us to predict 7 additional binding sites for LacI TFs, and to strengthen our confidence in two of the previously identified binding sites.

2.4. Experimental validation of new TF-TBS couples

In total 41 LacI-TFs were selected for protein-DNA interaction study by EMSAs. Proteins were assessed for their production levels in different cultures conditions (temperature, incubation time post induction) in order to choose one where a majority of them were produced. Their solubility, purification degree, and their stability as pure proteins after mid- or long-term storage at −20 °C were also assessed, after purification. According to these criteria, 16 6His-tagged LacI-TFs were retained for EMSAs (Fig. 3). DNA probes containing the MEME predicted binding site and tagged with Cy5 were incubated with increasing concentrations of their respective purified LacI-TFs as described previously [47, 54]. DNA-protein interactions were observed using an ImageQuant™ LAS 4000, by detecting the fluorescence emission of the Cy5-tag using a 670 nm detection filter. ManR (LacI 003, Fig. 3 second panel) was used as a positive control for the EMSA method, as its cognate palindromic motif GACAACGTTGTC has been recently confirmed experimentally [41]. Interestingly, no retardation could be observed for LacI 001 (panel 1 in Fig. 3), whose binding site is classified in the B category because of the lack of a central CG-pair. For the remaining 14 tested TF-TFBS couples a retardation band could be observed. The high success rate of the DNA-protein interaction assays demonstrates that the AURTHO approach is an appropriate way of discovering highly reliable TFBSs for unstudied TFs.

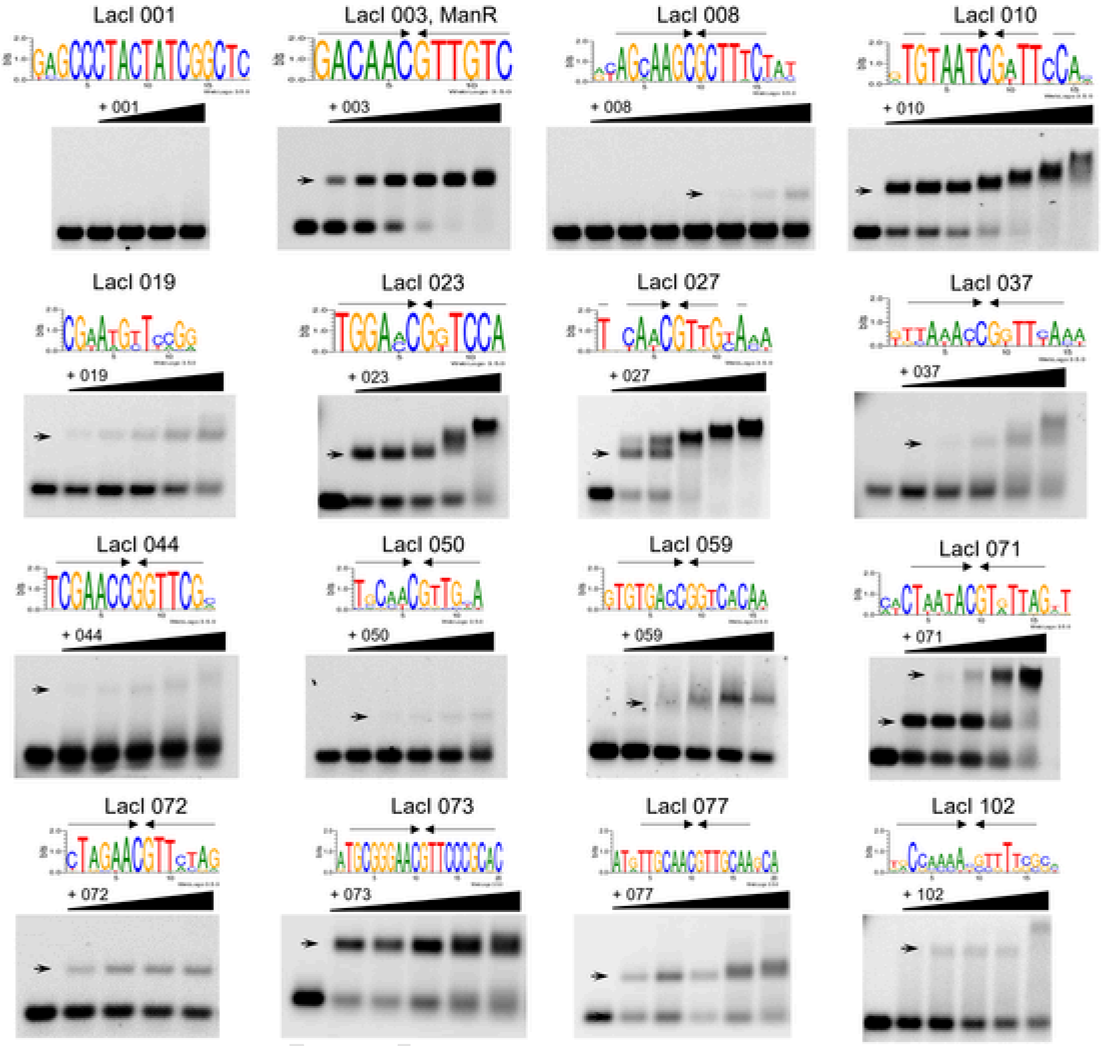


Fig. 3. Experimental validation of a selection of LacI TF-TFBS couples by EMSAs. Arrows above the WebLogos indicate the position of the inverted repeat. The arrow in the gels points to the first condition where a retarded band was observed. The ManR probe was used as positive control for a TF-TFBS couple previously already validated experimentally [41] Note the absence of retardation for LacI001 which does not possess the central CG pair, nor presents a symmetrical dyad. We used two-fold serial dilutions in order to create a range of concentrations of the pure protein for the EMSAs.

3. Conclusions and perspectives

Identifying the DNA sequence bound by a TF is key to unveiling novel regulatory pathways and attributing novel biological functions to genes/proteins that belong to a regulon. In this work, we assessed to which extent a *de novo* methodology based on the assumption that a large proportion of TFs control their own expression would be able to provide a reliable candidate TFBS for a TF with unknown function. The AURTHO approach drastically narrows down the searched regions for TFBSs in the bacterial chromosome, mainly focusing the DNA motif enrichment analysis within the upstream region of the TF of interest. Using TFs member of the LacI family in the *Streptomyces* genus as a case study, we identified 88 highly reliable TFBSs that possess the hallmarks of most LacI family regulator binding sites, *i.e.*, a CG pair centred in a symmetric dyad. All the DNA probes tested containing a motif with these sequence characteristics showed positive and specific interaction through EMSAs with their associated pure LacI TF, thereby demonstrating the high reliability of the predicted TFBSs. Hence, our approach showcases a very high potential at revealing the DNA sequences bound by a transcriptional regulator, as before our work, about four decades of study managed to reveal the TFBS of only 5 LacI-family TFs in *Streptomyces* species. This represents a potential improvement of 18-fold compared to the current state of knowledge. The main limitation resides on the number of members within a COG which directly affects the number of upstream regions to align for finding a conserved motif. When we initiated this work in 2018, 90 *Streptomyces* complete genomes were available and from these data, 53 motifs were predicted from 172 COGs (orphan COGs included). Little more than a year later (October 2019), the number of complete genomes from this genus had roughly doubled (182 genomes, this work), and the AURTHO methodology yielded 90 motifs for 167 COGs (orphan COGs included). As the number of COGs negligibly changed (~3 %) between both analyses while the number of motifs found almost doubled, this considerable improvement has to be imputed to the substantial portion of COGs that were not orphan anymore which allowed our methodology to be applicable. This reflects that the successive rounds of the AURTHO approach will become more and more successful at predicting putative *cis*-elements as the number of available genomes of one taxon increases.

One crucial question when applying phylogenetic footprinting, is the choice of the phylogenetic distance between the taxa selected for analysis. Indeed, the analyzed species can be neither too closely related (too much conservation in the regulatory region, alignment uninformative), nor too distant (the regulatory element will not be conserved). We show that, when a study is focused on a specific bacterial genus, the AURTHO approach is very potent at highlighting taxon-specific regulatory interactions, compared to the ones available in the RegPrecise database. In the latter, only 11 LacI TFs in the *Streptomyces* genus have been highlighted through regulon reconstruction and propagation, all of which are highly conserved and probably have orthologs in other genera. As shown in Fig. 4, for 10 of them, the motifs proposed by both approaches were either identical or highly similar. The remarkable exception relates to the second most conserved *Streptomyces* LacI-COG (002, with SCO4158 as the representative member). Indeed, the RegPrecise motif for this regulator is a palindromic and CG centred sequence (TCTACGCGCGTAGA), while our predicted motif (CGCGTAGACT) partially corresponds to the half right part of the palindrome, the other half being degenerated and not conserved (Fig. 4). The possible lack of autoregulation of LacI002 raises the question of whether this regulator is a global one (high number of target genes whose functions pertain to different cellular processes), as it has been suggested that global LacI TFs are less likely (~50 %) to use an autoregulatory mechanism, compared to local regulators (~75 %) [28]. And indeed, preliminary regulon identification revealed that the scope of the regulatory action of SCO4158 is extensive. Therefore, identifying the TFBS of global regulators *via* an analysis of their upstream region could be relatively less successful at providing reliable candidate motifs. This result suggests that both “AURTHO” and “regulon propagation and reconstruction” approaches are complementary, the latter being more adequate when focusing on global regulators with conserved regulatory interactions across a more phylogenetically diverse group.

Past studies on the conservation of TF-TFBS couples in distant bacterial groups suggest that the AURTHO approach will also generate a similar rate of success/reliability when applied to orthologs that do not belong to a same/unique genus [55,56]. Additionally, autoregulation has also been frequently observed for TF belonging to other families, such as GntR, MerR, MarR, IclR, among many others. Some binding sites from these families have also been characterized, hence using the hallmarks of these TFBSs combined with the AURTHO approach will surely increase the discovery rate of novel conserved motifs in these families as well. Overall, the results presented in this work suggest that the AURTHO approach will greatly facilitate the discovery of a plethora of *cis*-acting elements in all bacterial genus.

4. Materials and methods

4.1. Bioinformatics

Genome assemblies belonging to the genus *Streptomyces* were downloaded from the NCBI database and filtered based on the “Complete Genome” (assembly_level) and “latest” (version_status) tags in the assembly summary file. Proteinortho (v6.0.8) was used to compare all protein sequences and cluster them into orthologous groups (COGs). This version uses diamond (v0.9.36) as a default sequence aligner, and clusters groups based on the reciprocal best alignment heuristic (RBAH) [32]. HMMER3 was used to perform hmmscan on all proteins and identify protein domains by comparing them to the domain profiles in the Pfam-A database [33–35]. The P2TF database was used as a guide for TF identification based on the proteins' domain combinations [36]. The MEME software (Multiple Em for Motif Elicitation, v5.1.0) was used to align upstream regions of identified LacI TF genes and identify putative transcription factor binding sites [37,57]. Based on the previously described LacI TFBS hallmarks, the most probable motif(s) were selected and downloaded in FASTA format, and a sequence logo was created with WebLogo3 (WebLogo v3.5.0) [38]. Position Weight Matrices (PWM) were calculated on R using the Biostrings package (https://bioconductor.org/packages/Biostrings) and expressed as a log-

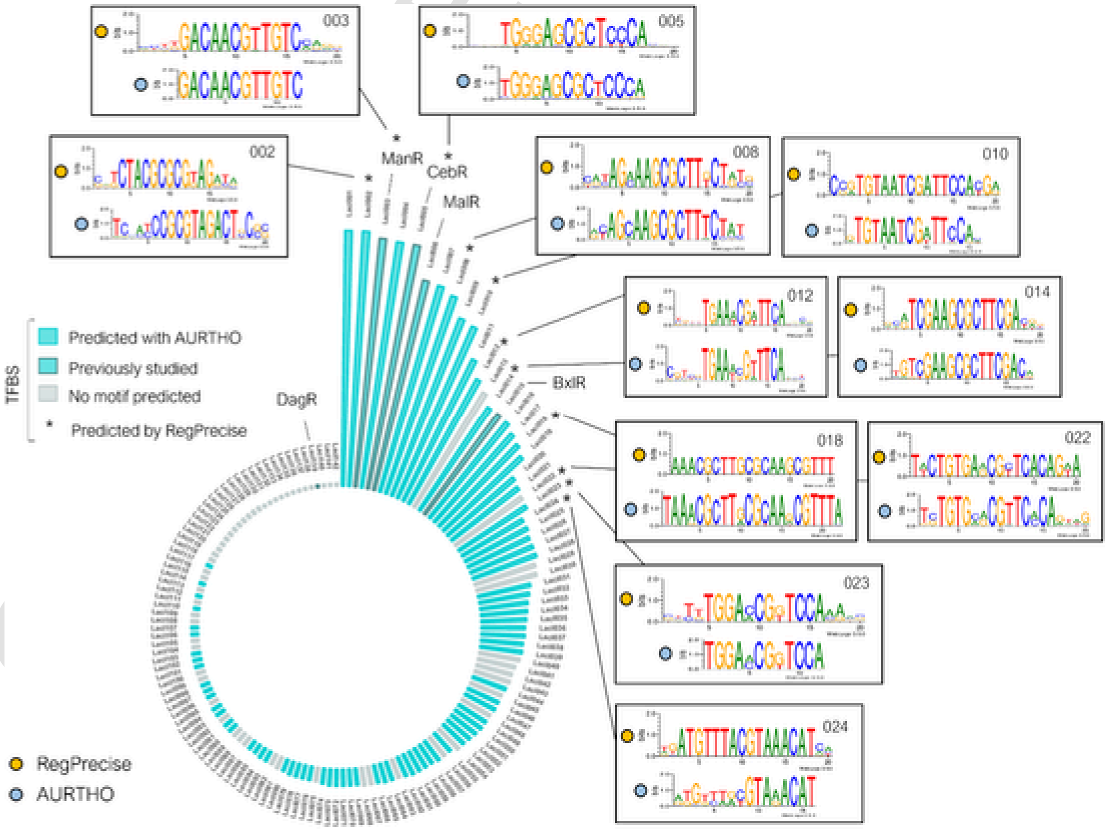


Fig. 4. TFBSs predicted by the AURTHO methodology and comparison with TFBSs available in the RegPrecise database. The height of the bars represents the number of genes contained in each COG. Bars colored in blue indicate the motif for the COG was predicted using the AURTHO approach, while bars colored in grey indicate COGs for which no motif could be predicted. For 11 COGs, an additional box is linked to the COG, containing the motif predicted by RegPrecise (yellow circle) and the motif predicted by AURTHO (blue circle). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

likelihood [23]. The PWMs calculated with different background nucleotide probabilities (reflecting either a 50 % or a 71.3 % GC content, the latter being the average GC content in the *Streptomyces* genus) are available in Supplementary files (pwm50.tar.gz and pwm71.tar.gz).

4.2. Heterologous production and purification of His-tagged proteins

The 41 LacI-TF genes selected for DNA-protein interaction studies are listed in Table S1. 40 of them were ordered at Twist Biosciences for codon-optimized sequence cloned in the *Nde*I and *Xho*I restriction sites of pET-28a for heterologous production in *E. coli* BL21(DE3). In addition, we used pSIN002 in which the original sequence of SCO1078 was cloned into the pET-22b (between NdeI and *Hin*dIII restriction sites) and was heterologously produced in the BL21 Rosetta™ (DE3) strain of *E. coli*. All proteins were 6His-tagged on their C-terminal extremity, enabling Immobilised Metal Affinity Chromatography (IMAC) purification on an Ni-NTA column from Cytiva (HisTrap™ HP). Transformed *E. coli* strains were inoculated in TB (Terrific Broth) supplemented with the appropriate antibiotics for plasmid selection (kanamycin for pET-28a, ampicillin and chloramphenicol for pET-22b and pLysS-containing *E. coli* Rosetta™ strains). The production was induced with 1 mM of IPTG when the culture attained an optical density of 0.8 (at 600 nm), and the culture was left overnight at 37 °C. The next day, pelleted cells (10.000 rpm, 30 min, 4 °C) were resuspended in 50 mL of Equilibration buffer (see below for composition) and lysed using a high-pressure homogeniser (Avestin Emulsiflex C3). After another round of centrifugation (18.000 rpm, 30 min, 4 °C), the supernatant, corresponding to the soluble intracellular fraction of the lysis mixture was filtered (0.22 μM) before IMAC purification. Buffers used for the protein purification process were of the following composition: (i) equilibration buffer (50 mM Phosphate Buffer, 20 mM imidazole, 1 M NaCl, pH 7.5), (ii) wash buffer, (50 mM Phosphate Buffer, 20 mM imidazole, 2 M NaCl, pH 7.5), (iii) elution buffer (50 mM Phosphate Buffer, 500 mM imidazole, 150 mM NaCl, pH 7.5). The protein purification was performed on the NGC Quest 10 Chromatography and the NGC Quest 100 Chromatography (Bio-Rad) at the Protein Factory platform (InBioS-CIP, ULiège). Selected fractions based on the absorbance at 280 nm of the elution profile were deposited on SDS-PAGE gels (Mini-PROTEAN® TGX™ Precast Gels, Bio-Rad) gels to assess their purity, and the most concentrated ones were desalted using a HiPrep™ 26/10 desalting column (packed with Sephadex® G-25 Fine) from Cytiva. The resulting desalted fractions in EMSA buffer (Tris 10 mM pH 7.5, KCl 50 mM, DTT 1 mM, glycerol 2 %, $CaCl_2$ 0.25 mM, $MgCl_2$ 0.5 mM), were analyzed on SDS-PAGE gel (Mini-PROTEAN® TGX™ Precast Gels, Bio-Rad) for purity, and only the most concentrated and pure fractions were collected and used for DNA-protein interaction studies.

4.3. Electrophoretic mobility shift assays

DNA probes were designed using the predicted binding sites for each of the selected LacI COGs. For each COG, a matrix of possible binding sites (in FASTA format) was downloaded from MEME, and then used to create a WebLogo based on which we deduced the consensus sequence for designing the probe. In cases where a nucleotide was not overrepresented at a specific position, we chose the nucleotide complementary to the nucleotide conserved in the other part of the motif in order to make it closer to a dyad symmetry. The primers (Eurogentec, Seraing, Belgium) used to generate the DNA probes are listed in Supplementary Table S2. The interaction reactions between pure 6His-tagged proteins and their Cy5-labelled DNA probe containing their predicted binding site were performed in EMSA buffer (Tris 10 mM pH 7.5, KCl 50 mM, DTT 1 mM, glycerol 2 %, $CaCl_2$ 0.25 mM, $MgCl_2$ 0.5 mM), as described previously [47,54]. The final EMSA samples which were incubated at room temperature for 15 min contained 12.5 nM of hybridized probe, 1.5 mM of non-specific protein (Bovine Serum Albumin, BSA), 10 mg of non-specific DNA (sheared Salmon Sperm DNA, Invitrogen™), representing a 400-fold excess compared to the probe, and increasing concentrations of protein (obtained by performing two-fold serial dilutions of the fraction with the highest concentration of protein). After migration into a 1 % agarose gel, the visualization of the free and retarded bands was monitored using the fluorescence imager (GE Healthcare), detecting the Cy5-tagged DNA probes at a wavelength of 670 nm.

CRediT authorship contribution statement

Sinaeda Anderssen : Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Aymeric Naômé :** Conceptualization, Methodology, Validation, Formal analysis, Data curation, Writing – review & editing. **Cédric Jadot :** Investigation, Writing – review & editing. **Alain Brans :** Conceptualization, Writing – review & editing. **Pierre Tocquin :** Conceptualization, Methodology, Resources, Writing – review & editing, Funding acquisition. **Sébastien Rigali :** Conceptualization, Validation, Writing – original draft, Writing – review & editing, Visualization, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Sebastien Rigali reports a relationship with Hedera 22 that includes: board membership and consulting or advisory. Aymeric Naome reports a relationship with Hedera 22 that includes: employment. Pierre Tocquin reports a relationship with Hedera 22 that includes: board membership, consulting or advisory, and non-financial support.

Data availability

All in-house scripts that were used to generate the data are available on GitHub (https://github.com/SinaedaA/AURTHO), as well as a markdown file retracing all steps of the AURTHO methodology.

Acknowledgments

We are grateful to the teams working at the Protein Factory platform (https://www.proteinfactory.uliege.be/cms/c_14301576/en/proteinfactory) and the Robotein platform (https://www.robotein.uliege.be/cms/c_14301428/en/robotein) for productive discussions and assistance in experimental design. S.R. is a Fonds de la Recherche Scientifique (FRS-FNRS) senior research associate.

Data availability

All in-house scripts that were used to generate the data (genome download, COG creation, TF family identification, upstream sequence extraction, MEME analysis) are available on GitHub (https://github.com/SinaedaA/AURTHO), as well as a markdown file retracing all steps of the AURTHO methodology.

Funding

This work was supported by ‘Fonds De La Recherche Scientifique - FNRS’ [FRIA 1.E.031.18-20 to SA and SR, R.FNRS.5240 to SR]; and the ‘Gouvernement Wallon’ [1510530 to AN and SR].

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.bbagrm.2022.194847.

References

[1] D.F. Browning, M. Butala, S.J.W. Busby, Bacterial transcription factors: regulation by pick "N" mix, J. Mol. Biol. 431 (20) (2019 Sep 20) 4067–4077.

[2] D.F. Browning, S.J.W. Busby, Local and global regulation of transcription initiation in bacteria, Nat. Rev. Microbiol. 14 (10) (2016 Oct) 638–650.

[3] C. Mejía-Almonte, S.J.W. Busby, J.T. Wade, J. van Helden, A.P. Arkin, G.D. Stormo, et al., Redefining fundamental concepts of transcription initiation in bacteria, Nat. Rev. Genet. 21 (11) (2020 Nov) 699–714.

[4] S.A.F.T. Van Hijum, M.H. Medema, O.P. Kuipers, Mechanisms and evolution of control logic in prokaryotic transcriptional regulation, Microbiol. Mol. Biol. Rev. 73 (3) (2009) 481–509.

[5] L.A. Baumgart, J.E. Lee, A. Salamov, D.J. Dilworth, H. Na, M. Mingay, et al., Persistence and plasticity in bacterial gene regulation, Nat. Methods 18 (12) (2021 Dec) 1499–1505.

[6] A. Santos-Zavaleta, H. Salgado, S. Gama-Castro, M. Sánchez-Pérez, L. Gómez-Romero, D. Ledezma-Tejeida, et al., RegulonDB v 10.5: tackling challenges to unify classic and high throughput knowledge of gene regulation in E. coli K-12, Nucleic Acids Res. 47 (D1) (2019 Jan 8) D212–D220.

[7] A. Bartlett, R.C. O'Malley, S.C. Huang, M. Galli, J.R. Nery, A. Gallavotti, et al., Mapping genome-wide transcription-factor binding sites using DAP-seq, Nat. Protoc. 12 (8) (2017) 1659–1672.

[8] A. Ishihama, T. Shimada, Y. Yamazaki, Transcription profile of Escherichia coli: genomic SELEX search for regulatory targets of transcription factors, Nucleic Acids Res. 44 (5) (2016 Mar 18) 2058–2074.

[9] B. Liu, J. Yang, Y. Li, A. McDermaid, Q. Ma, An algorithmic perspective of de novo cis-regulatory motif finding based on ChIP-seq data, Brief. Bioinform. 19 (5) (2018 Sep 28) 1069–1081.

[10] P.J. Park, ChIP–seq: advantages and challenges of a maturing technology, Nat. Rev. Genet. 10 (10) (2009 Oct) 669–680.

[11] Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics, Nat. Rev. Genet. 10 (1) (2009 Jan) 57–63.

[12] M.A. Świątek-Połatyńska, G. Bucca, E. Laing, J. Gubbens, F. Titgemeyer, C.P. Smith, et al., Genome-wide analysis of in vivo binding of the master regulator DasR in Streptomyces coelicolor identifies novel non-canonical targets, PLoS ONE 10 (4) (2015 Apr 15) e0122479.

[13] S. Dwarakanath, A.K. Chaplin, M.A. Hough, S. Rigali, E. Vijgenboom, J.A.R. Worrall, Response to copper stress in Streptomyces lividans extends beyond genes under direct control of a copper-sensitive operon repressor protein (CsoR) *, J. Biol. Chem. 287 (21) (2012 May 18) 17833–17847.

[14] C. Liao, S. Rigali, C.L. Cassani, E. Marcellin, L.K. Nielsen, B.-C. Ye, Control of chitin and N-acetylglucosamine utilization in Saccharopolyspora erythraea, Microbiology 160 (9) (2014 Sep 1) 1914–1928.

[15] S. Rigali, M. Schlicht, P. Hoskisson, H. Nothaft, M. Merzbacher, B. Joris, et al., Extending the classification of bacterial transcription factors beyond the helix-turn-helix motif as an alternative approach to discover new cis/trans relationships, Nucleic Acids Res. 32 (11) (2004) 3418–3426.

[16] D.A. Rodionov, Comparative genomic reconstruction of transcriptional regulatory networks in bacteria [Internet]. Chem. Rev. 107 (8) (2007) Available from: https://pubs.acs.org/sharingguidelines.

[17] L.-L. Yao, C.-H. Liao, G. Huang, Y. Zhou, S. Rigali, B. Zhang, et al., GlnR-mediated regulation of nitrogen metabolism in the actinomycete Saccharopolyspora erythraea, Appl. Microbiol. Biotechnol. 98 (18) (2014 Sep) 7935–7948.

[18] H. Li, V. Rhodius, C. Gross, E.D. Siggia, Identification of the binding sites of regulatory proteins in bacterial genomes, Proc. Natl. Acad. Sci. U. S. A. 99 (18) (2002 Sep 3) 11772–11777.

[19] M.M. Mwangi, E.D. Siggia, Genome wide identification of regulatory motifs in Bacillus subtilis, BMC Bioinform. 16 (4) (2003 May) 18.

[20] D.J. Studholme, S.D. Bentley, J. Kormanec, Bioinformatic identification of novel regulatory DNA sequence motifs in Streptomyces coelicolor, BMC Microbiol. 4 (1) (2004 Apr 8) 14.

[21] S. Rigali, A. Derouaux, F. Giannotta, J. Dusart, Subdivision of the helix-turn-helix GntR family of bacterial regulators in the FadR, HutC, MocR, and YtrA subfamilies, J. Biol. Chem. 277 (15) (2002) 12507–12515.

[22] R. Janky, J. van Helden, Evaluation of phylogenetic footprint discovery for predicting bacterial cis-regulatory elements and revealing their evolution, BMC Bioinforma. 9 (1) (2008 Jan 23) 37.

[23] W.W. Wasserman, A. Sandelin, Applied bioinformatics for the identification of regulatory elements, Nat. Rev. Genet. 5 (4) (2004 Apr) 276–287.

[24] M.D. Kazanov, X. Li, M.S. Gelfand, A.L. Osterman, D.A. Rodionov, Functional diversification of ROK-family transcriptional regulators of sugar catabolism in the Thermotogae phylum, Nucleic Acids Res. 41 (2) (2013 Jan) 790–803.

[25] S.A. Leyn, I.A. Suvorova, A.E. Kazakov, D.A. Ravcheev, V.V. Stepanova, P.S. Novichkov, et al., Comparative genomics and evolution of transcriptional regulons in Proteobacteria, Microb. Genomics 2 (7) (2016 Jul 11).

[26] P.S. Novichkov, A.E. Kazakov, D.A. Ravcheev, S.A. Leyn, G.Y. Kovaleva, R.A. Sutormin, et al., RegPrecise 3.0 - a resource for genome-scale exploration of transcriptional regulation in bacteria, BMC Genomics 14 (2013) 1.

[27] P.S. Novichkov, O.N. Laikova, E.S. Novichkova, M.S. Gelfand, A.P. Arkin, I. Dubchak, et al., RegPrecise: a database of curated genomic inferences of transcriptional regulatory interactions in prokaryotes, Nucleic Acids Res. 38 (Database issue) (2010 Jan) D111–D118.

[28] D.A. Ravcheev, M.S. Khoroshkin, O.N. Laikova, O.V. Tsoy, N.V. Sernova, S.A. Petrova, et al., Comparative genomics and evolution of regulons of the LacI-family transcription factors, Front. Microbiol. 5 (2014) 294.

[29] S.D. Bentley, K.F. Chater, A.-M. Cerdeño-Tárraga, G.L. Challis, N.R. Thomson, K.D. James, et al., Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2), Nature 417 (6885) (2002 May 9) 141–147.

[30] H. Tsujibo, M. Kosaka, S. Ikenishi, T. Sato, K. Miyamoto, Y. Inamori, Molecular characterization of a high-affinity xylobiose transporter of Streptomyces thermoviolaceus OPC-520 and its transcriptional regulation, J. Bacteriol. 186 (4) (2004 Feb 15) 1029–1037.

[31] A. Schlösser, A. Weber, H. Schrempf, K. M, B. B, B. W, et al., Synthesis of the Streptomyces lividans maltodextrin ABC transporter depends on the presence of the regulator MalR, FEMS Microbiol. Lett. 196 (1) (2001 Mar 1) 77–83.

[32] M. Lechner, S. Findeiß, L. Steiner, M. Marz, P.F. Stadler, S.J. Prohaska, Proteinortho: detection of (Co-)orthologs in large-scale analysis, 2011.

[33] S.R. Eddy, Accelerated profile HMM searches, PLoS Comput. Biol. 7 (10) (2011 Oct 20) e1002195.

[34] S. El-Gebali, J. Mistry, A. Bateman, S.R. Eddy, A. Luciani, S.C. Potter, et al., The Pfam protein families database in 2019, Nucleic Acids Res. 47 (D1) (2019 Jan 8) D427–D432.

[35] J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G.A. Salazar, E.L.L. Sonnhammer, et al., Pfam: the protein families database in 2021, Nucleic Acids Res. 49 (D1) (2021 Jan 8) D412–D419.

[36] P. Ortet, G. De Luca, D.E. Whitworth, M. Barakat, P2TF: a comprehensive resource for analysis of prokaryotic transcription factors [cited 2017 Aug 23]. BMC Microbiol. 13 (628) (2012) Available from: http://www.p2tf.org/.

[37] T.L. Bailey, C. Elkan, in: Fitting a mixture model by expectation maximization to discover motifs in biopolymers, 1994, pp. 28–36.

[38] G.E. Crooks, G. Hon, J.-M. Chandonia, S.E. Brenner, WebLogo: a sequence logo generator, Genome Res. 14 (2004) 1188–1190.

[39] D.A. Hodgson, Primary metabolism and its control in streptomycetes: a most unusual group of bacteria [cited 2022 Mar 29], in: Advances in Microbial Physiology, Academic Press, 2000, pp. 47–238 Available from: https://www.sciencedirect.com/science/article/pii/S0065291100420035.

[40] A. van der Meij, S.F. Worsley, M.I. Hutchings, G.P. van Wezel, Chemical ecology of antibiotic production by actinomycetes, FEMS Microbiol. Rev. 41 (3) (2017 May 1) 392–416.

[41] K. Ohashi, S. Hataya, A. Nakata, K. Matsumoto, N. Kato, W. Sato, et al., Mannose- and mannobiose-specific responses of the insect-associated cellulolytic bacterium Streptomyces sp.sStrain SirexAA-E, Appl. Environ. Microbiol. 87 (14) (2021) e02719–e02720.

[42] J. Nguyen, The regulatory protein Reg1 of Streptomyces lividans binds the promoter region of several genes repressed by glucose, FEMS Microbiol. Lett. 175 (1) (1999 Jun) 51–58.

[43] J. Nguyen, F. Francou, M.J. Virolle, M. Guérineau, Amylase and chitinase genes in Streptomyces lividans are regulated by reg1, a pleiotropic regulatory gene, J. Bacteriol. 179 (20) (1997 Oct) 6383–6390.

[44] G.P. van Wezel, J. White, P. Young, P.W. Postma, M.J. Bibb, Substrate induction and glucose repression of maltose utilization by Streptomyces coelicolor A3(2) is controlled by malR, a member of the lacI-galR family of regulatory genes, Mol. Microbiol. 23 (3) (1997 Jan 1) 537–549.

[45] G.P. van Wezel, J. White, M.J. Bibb, P.W. Postma, The malEFG gene cluster of Streptomyces coelicolor A3(2): characterization, disruption and transcriptional analysis, Mol Gen Genet. 254 (5) (1997 May 1) 604–608.

[46] A.J. Book, G.R. Lewin, B.R. Mcdonald, T.E. Takasuka, E. Wendt-Pienkowski, D.T. Doering, et al., Evolution of high cellulolytic activity in symbiotic streptomyces through selection of expanded gene content and coordinated gene expression [cited 2017 Aug 24]. PLoS Biol. 14 (6) (2016) Available from: http://journals.plos.org/plosbiology/article/file?id = 10.1371/journal.pbio.1002475&type = printable.

[47] I.M. Francis, S. Jourdan, S. Fanara, R. Loria, S. Rigali, The cellobiose sensor CebR is the gatekeeper of Streptomyces scabies pathogenicity, mBio 6 (2) (2015 Feb 24) e02018.

[48] S. Jourdan, I.M. Francis, M.J. Kim, J.J.C. Salazar, S. Planckaert, J.-M. Frère, et al., The CebE/MsiK transporter is a doorway to the cello-oligosaccharide-mediated induction of Streptomyces scabies pathogenicity, Sci. Rep. 6 (January) (2016) 27144.

[49] K. Marushima, Y. Ohnishi, S. Horinouchi, CebR as a master regulator for cellulose/cellooligosaccharide catabolism affects morphological development in Streptomyces griseus, J. Bacteriol. 191 (19) (2009 Oct 1) 5930–5940.

[50] A. Schlösser, T. Aldekamp, H. Schrempf, R.G. B, H.C. P, M.A. S, et al., Binding characteristics of CebR, the regulator of the ceb operon required for cellobiose/cellotriose uptake in Streptomyces reticuli, FEMS Microbiol. Lett. 190 (1) (2000 Sep 1) 127–132.

[51] F. Giannotta, S. Rigali, M.-J. Virolle, J. Dusart, Site-directed mutagenesis of conserved inverted repeat sequences in the xylanase C promoter region from Streptomyces sp.EC3, Mol. Genet. Genomics (270) (2003) 337–346.

[52] F. Giannotta, J. Georis, A. Moreau, C. Mazy-Servais, B. Joris, J. Dusart, A sequence-specific DNA-binding protein interacts with the xZnC upstream region of Streptomyces sp. strain EC3, FEMS Microbiol. Lett. 142 (1996) 91–97.

[53] M. Tsevelkhoroloo, S.H. Shim, C.-R. Lee, S.-K. Hong, Y.-S. Hong, LacI-family transcriptional regulator DagR acts as a repressor of the agarolytic pathway genes in Streptomyces coelicolor A3(2), Front. Microbiol. 6 (12) (2021 Apr) 658657.

[54] E. Tenconi, M. Urem, M.A. Świątek-Połatyńska, F. Titgemeyer, Y.A. Muller, G.P. van Wezel, et al., Multiple allosteric effectors control the affinity of DasR for its target sites, Biochem. Biophys. Res. Commun. 464 (1) (2015 Aug 14) 324–329.
[55] R. Bertram, S. Rigali, N. Wood, A.T. Lulko, O.P. Kuipers, F. Titgemeyer, Regulon of the N-acetylglucosamine utilization regulator NagR in Bacillus subtilis, J. Bacteriol. 193 (14) (2011 Jul 15) 3525–3536.
[56] M. Urem, T. van Rossum, G. Bucca, G.F. Moolenaar, E. Laing, M.A. Świątek-Połatyńska, et al., OsdR of Streptomyces coelicolor and the dormancy regulator DevR of Mycobacterium tuberculosis control overlapping regulons, in: M. Traxler (Ed.), mSystems, 1(3), 2016, pp. e00014–e00016.
[57] T.L. Bailey, J. Johnson, C.E. Grant, W.S. Noble, The MEME suite, Nucleic Acids Res. 43 (W1) (2015 Jul 1) W39–W49.