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

Sinaeda Anderssen[1], Aymeric Naômé[1, 2], Cédric Jadot[1], Alain Brans[1], Pierre Tocquin[2,3], and Sébastien Rigali[1, 2, *]

[1]InBioS – Center for Protein Engineering, University of Liège, B-4000 Liège, Belgium
[2]HEDERA 22, Boulevard du Rectorat 27b, B-4000 Liège, Belgium
[3]InBioS – PhytoSystems, University of Liège, B-4000 Liège, Belgium
*Corresponding author: Tel: +32 4 366 33 77; Email: srigali@uliege.be

Keywords: Autoregulation; Transcription factor binding site discovery; Regulon Prediction; *cis*-acting elements;

Key points:

1. Transcription factor (TF) autoregulation implies that their binding site (TFBS) is in their close vicinity
2. We developed and assessed the AURTHO methodology (AUtoregulation of oRTHOlogous TFs) for TFBS discovery
3. Our results shows that AURTHO greatly facilitates the identification of highly reliable novel TFBSs

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 neighbouring 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.

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 (Browning et al., 2019; Browning & Busby, 2016; Mejía-Almonte et al., 2020; Van Hijum et al., 2009). 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* (Baumgart et al., 2021; Santos-Zavaleta et al., 2019). 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 (Bartlett et al., 2017; Baumgart et al., 2021; Ishihama et al., 2016; Liu et al., 2018; Park, 2009; Wang et al., 2009). 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 (Świątek-Połatyńska et al., 2015). 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' functions (Dwarakanath et al., 2012; Liao et al., 2014; Rigali et al., 2004; Rodionov, 2007; Van Hijum et al., 2009; Yao et al., 2014). 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 over-represented dyad-type motifs in intergenic regions of the genome, where *cis*-acting elements are expected to be found (Li et al., 2002; Mwangi & Siggia, 2003; Studholme et al., 2004). 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 (Rigali et al., 2002, 2004). Now that the number of available genomes has significantly grown, approaches based on comparative genomics and more specifically on phylogenetic

footprinting, have become possible (Janky & van Helden, 2008; Rodionov, 2007; Wasserman & Sandelin, 2004). 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 (Kazanov et al., 2013; Leyn et al., 2016; Novichkov et al., 2010, 2013; Ravcheev et al., 2014; Rodionov, 2007).

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) (Bentley et al., 2002). 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.

RESULTS AND DISCUSSION

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 characterised by palindromic sequences of even length which contain a typical CG-pair at the centre of the motif (Ravcheev et al., 2014). Nonetheless, some atypical binding sites have been identified, showcasing uneven lengths, the absence of a CG-pair in the centre (Tsujibo et al., 2004), directed repeats (Schlösser et al., 2001) 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) (Ravcheev et al., 2014).

The methodology that guided our approach is detailed in the flowchart presented in Figure 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 (Lechner et al., 2011) 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 (Eddy, 2011; El-Gebali et al., 2019; Mistry et al., 2021), and TF genes were classified into families through signature domain combinations, as described in the P2TF database (Ortet et al., 2012). 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) (Ravcheev et al., 2014). 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 (Ortet et al., 2012). 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 (Bailey & Elkan, 1994) 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 (Ravcheev et al., 2014). Finally, the FASTA-format matrices of putative binding sites were used to create sequence logos with WebLogo (Crooks et al., 2004) 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 Figure 1 and Figure 2).

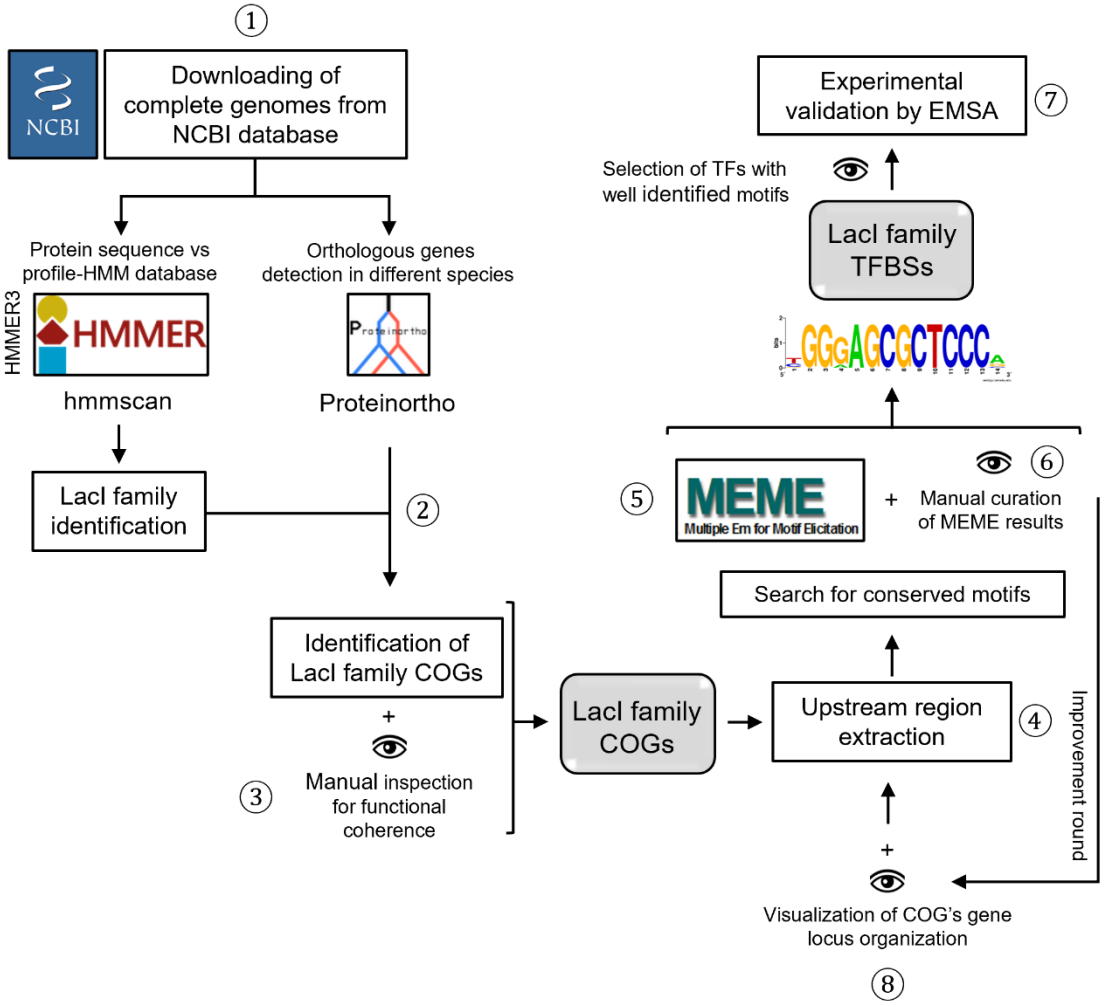


Figure 1. Flowchart illustrating the different steps of the AURTHO approach. The "eye" icon indicates the different steps that required manual inspection of the software/algorithm output before proceeding to the following step(s) of the methodology and/or to improve/increase the quality/quantity of the 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.

de novo identification of binding sites of LacI-family TFs in streptomycetes

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 (Hodgson, 2000;

van der Meij et al., 2017), LacI TFs were identified in all 182 studied complete genomes (supplementary Figure S1). However, there was a great disparity 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 (Figure 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) (Ohashi et al., 2021), (ii) the maltose/maltodextrin catabolism pathway regulator MalR (LacI005 in Table 1, conserved in 176/182 species) (Nguyen, 1999; Nguyen et al., 1997; Schlösser et al., 2001; van Wezel, White, Bibb, et al., 1997; van Wezel, White, Young, et al., 1997), iii) the cellulose/cello-oligosaccharide utilisation regulator CebR (LacI006 in Table 1, conserved in 153/182 species) (Book et al., 2016; Francis et al., 2015; Jourdan et al., 2016; Marushima et al., 2009; Schlösser et al., 2000), iv) the xylan/xylo-oligosaccharide utilization repressor BxlR (LacI015 in Table 1, conserved in 88/182 species) (Giannotta et al., 1996, 2003; Tsujibo et al., 2004), and v) the agar-utilisation regulator DagR (LacI139 in Table 1) (Tsevelkhoroloo et al., 2021), 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.

Identification of TF binding sites

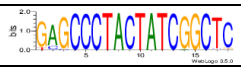
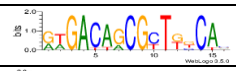
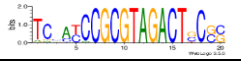
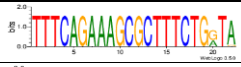
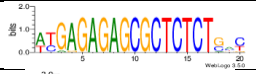
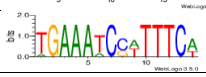
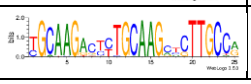
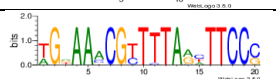
For each the 167 LacI-family COGs, a set of upstream regions was extracted with varying lengths as described in the Methodology section. This resulted in 138 sets of two or more upstream regions. Indeed, in the remaining cases, the COG was either orphan (one gene), or there was either only one, or no 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. This was supported by the previous observation of Ravcheev *et al* (Ravcheev et al., 2014) that LacI binding sites are rarely found beyond 300 nucleotides upstream of the target gene, or after the beginning of the coding region.

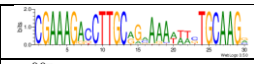
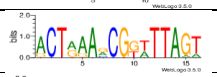
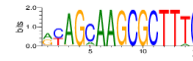
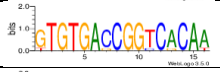
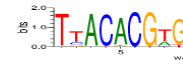
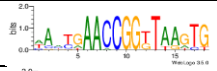
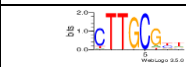
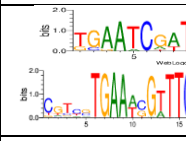
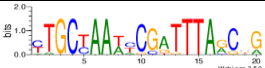
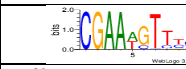
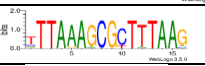
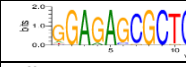
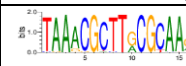
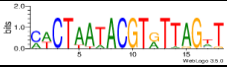
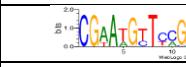
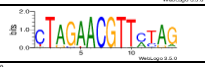
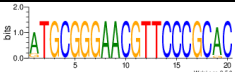
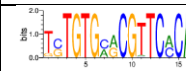
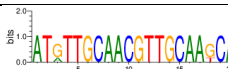
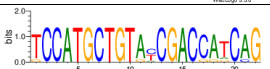
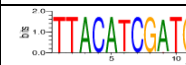
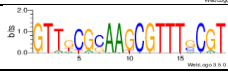
In order to first assess the reliability of our *de novo* approach, we singled out the studied LacI regulators (ManR, MalR, CebR, BxlR, and DagR), and checked if the motifs we generated using our *in-silico* approach correspond to their experimentally determined *cis*-acting sequences. As presented in Table 1, for ManR (LacI003), CebR (LacI006), and BxlR (LacI015), the *de novo* identified motifs were identical

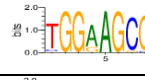
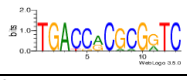
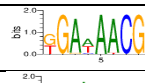
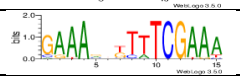
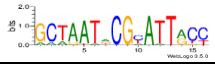
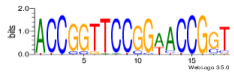
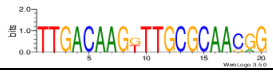
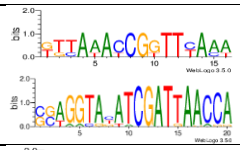
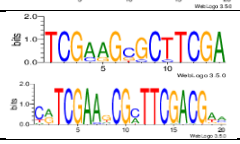
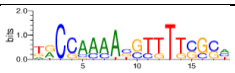
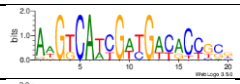
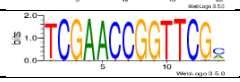
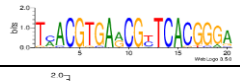
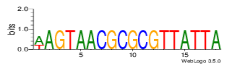
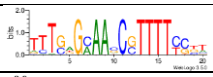
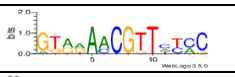
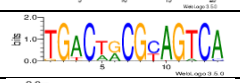
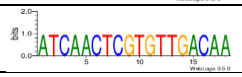
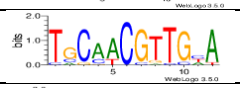
to their experimentally identified consensus sequences, i.e., GACAACGTTGTC (Ohashi et al., 2021), TGGGAGCGCTCCCA (Schlösser et al., 2000), and CGAA-Nx-TTCG (Giannotta et al., 1996, 2003; Tsujibo et al., 2004), respectively. For MalR (LacI005), the two binding sites deduced by DNase footprinting assays (Schlösser et al., 2001) 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 (Tsevelkhoroloo et al., 2021).

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 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).

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)
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 Figure 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, (Tsevelkhoroloo et al., 2021)).

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 Figure 2A). In most cases (LacI120 – LacI142 in Figure 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 (Figure 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. Figure 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 Figure 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) (Figure 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.

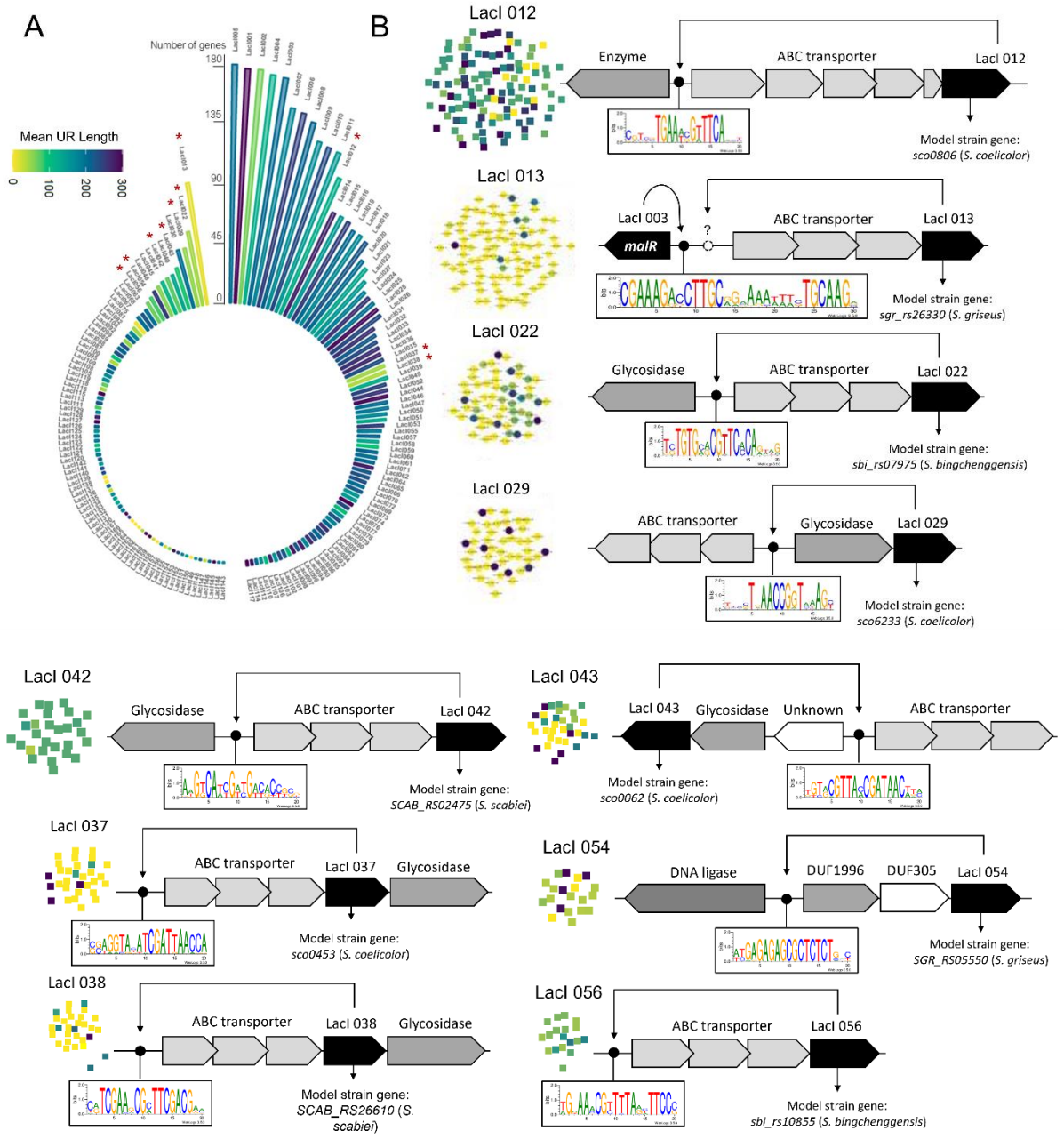


Figure 2. Inspection of the length of the TF upstream region and the gene locus organization. (**A**) Average length of upstream regions for each COG. The height of the bars is relative to the number of members in each COG, and the color indicates the mean length of extracted upstream regions. The right 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 analysed 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.

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 -20ºC were also assessed, after purification. According to these criteria, 16 6His-tagged LacI-TFs were retained for EMSAs (Figure 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 (Francis et al., 2015; Tenconi et al., 2015). 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, Figure 3 second panel) was used as a positive control for the EMSA method, as its cognate palindromic motif GACAACGTTGTC has been recently confirmed experimentally (Ohashi et al., 2021). Interestingly, no retardation could be observed for LacI 001 (panel 1 in Figure 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.

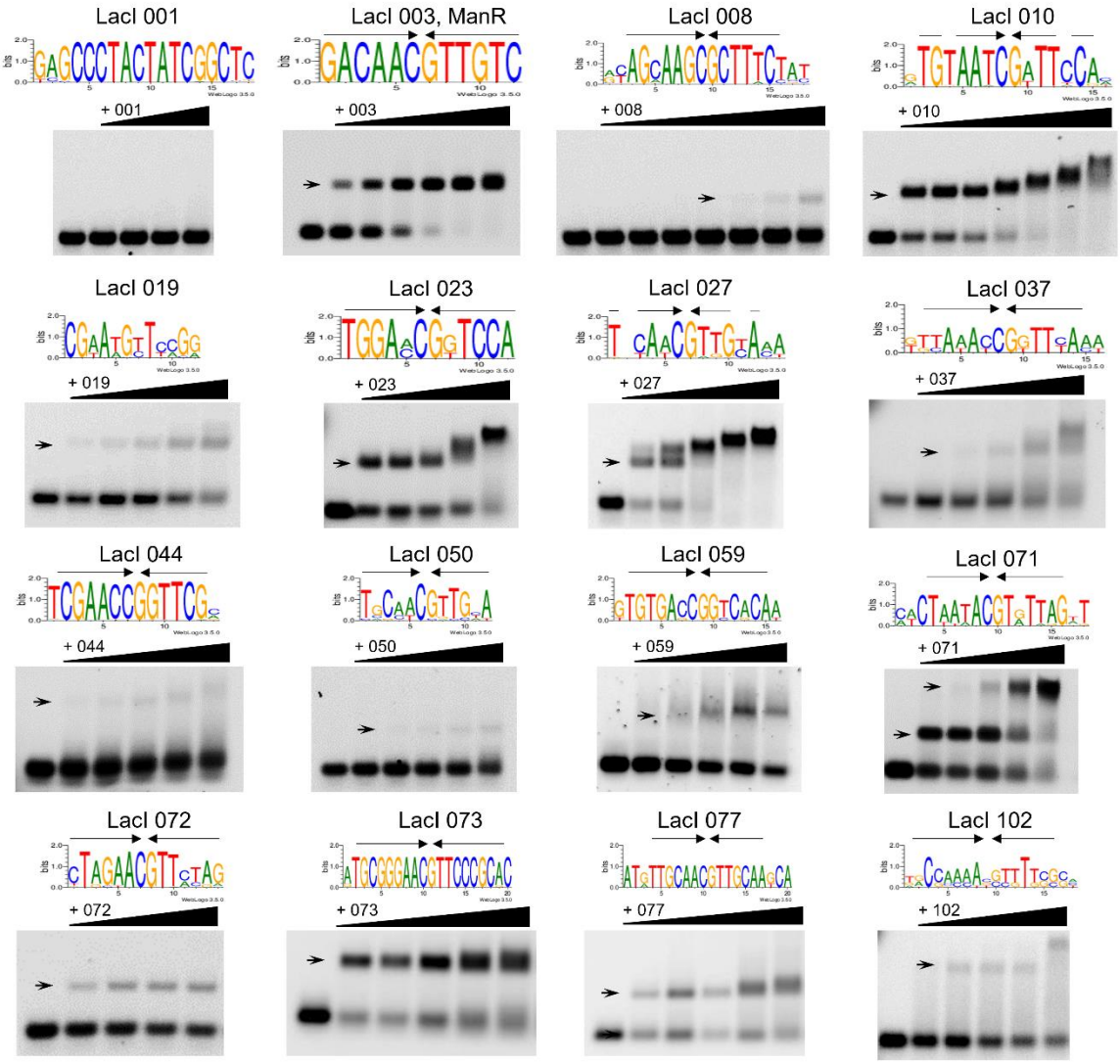


Figure 3. Experimental validation of a selection of LacI TF-TFBS couples by EMSAs. Arrows above the WebLogos indicate the position of the inverted repeat. The arrow in the gels points to the first condition where a retarded band was observed. The ManR probe was used as positive control for a TF-TFBS couple previously already validated experimentally (Ohashi et al., 2021) 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

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 centered 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.

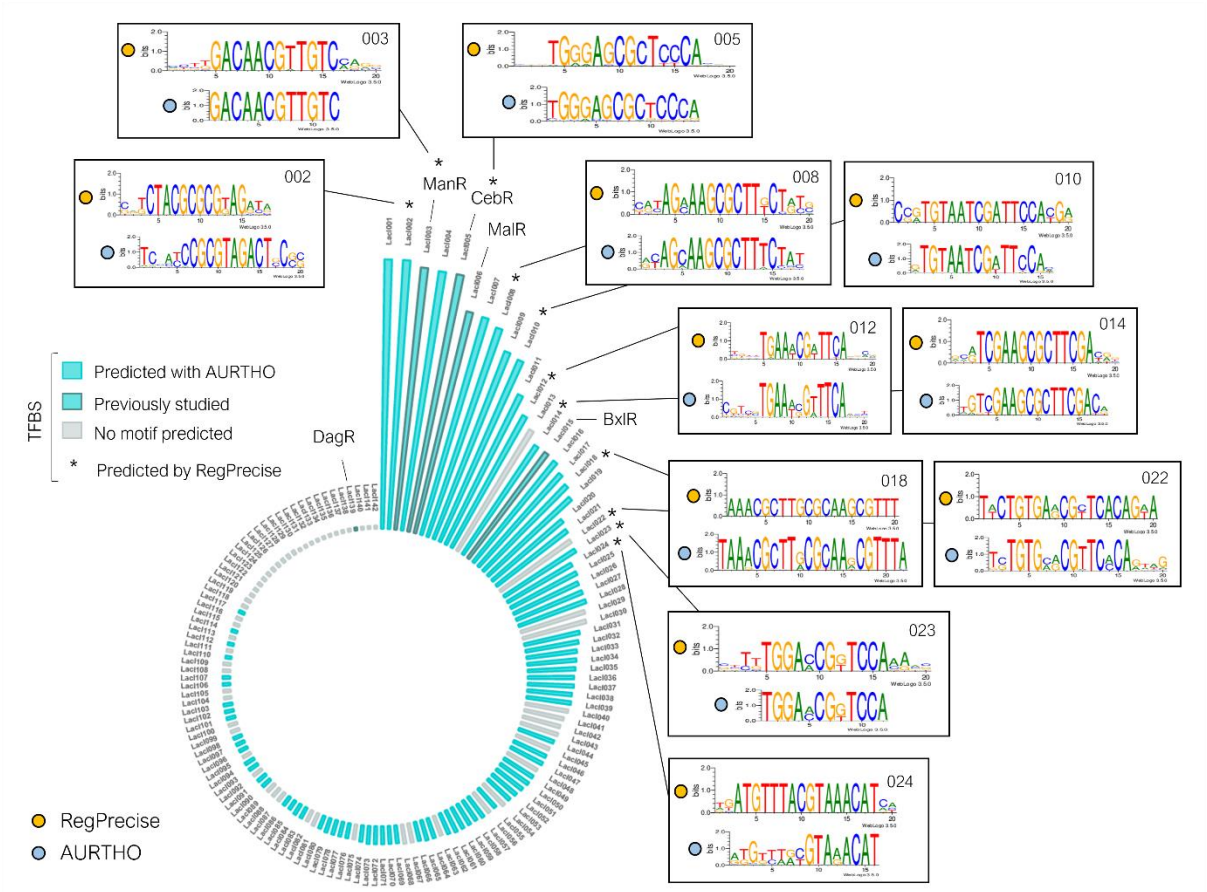


Figure 4. TFBSs predicted by the AURTHO methodology and comparison with TFBSs available in the RegPrecise database. The height of the bars represents the number of genes contained in each COG. Bars colored in blue indicate the motif for the COG was predicted using the AURTHO approach, while bars colored 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).

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 Figure 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 (Figure 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%) (Ravcheev et al., 2014). 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 orthologues that do not belong to a same/unique genus (Bertram et al., 2011; Urem et al., 2016). 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 suggests that the AURTHO approach will greatly facilitate the discovery of a plethora of *cis*-acting elements in all bacterial genus.

MATERIALS AND METHODS

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) (Lechner et al., 2011). 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 (Eddy, 2011; El-Gebali et al., 2019; Mistry et al., 2021). The P2TF database was used as a guide for TF identification based on the proteins' domain combinations (Ortet et al., 2012). 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 (Bailey et al., 2015; Bailey & Elkan, 1994). 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) (Crooks et al., 2004). Position Weight Matrices (PWM) were calculated on R using the Biostrings package (https://bioconductor.org/packages/Biostrings) and expressed as a log-likelihood (Wasserman & Sandelin, 2004). 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).

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 NdeI and XhoI 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 HindIII 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 37ºC. The next day, pelleted cells (10.000 rpm, 30 min, 4ºC) 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, 4ºC), 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, 1M 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 10mM pH 7.5, KCl 50mM, DTT 1mM, glycerol 2%, $CaCl_2$ 0.25 mM, $MgCl_2$ 0.5 mM), were analysed 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.

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 10mM pH 7.5, KCl 50mM, DTT 1mM, glycerol 2%, $CaCl_2$ 0.25 mM, $MgCl_2$ 0.5 mM), as described previously (Francis et al., 2015; Tenconi et al., 2015). 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 Albumine, 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.

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].

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.

REFERENCES

Bailey, T. L., & Elkan, C. (1994). *Fitting a mixture model by expectation maximization to discover motifs in biopolymers*. 28–36.

Bailey, T. L., Johnson, J., Grant, C. E., & Noble, W. S. (2015). The MEME Suite. *Nucleic Acids Research*, *43*(W1), W39–W49. https://doi.org/10.1093/nar/gkv416

Bartlett, A., O'Malley, R. C., Huang, S. C., Galli, M., Nery, J. R., Gallavotti, A., & Ecker, J. R. (2017). Mapping genome-wide transcription-factor binding sites using DAP-seq. *Nature Protocols*, *12*(8), 1659–1672. https://doi.org/10.1038/nprot.2017.055

Baumgart, L. A., Lee, J. E., Salamov, A., Dilworth, D. J., Na, H., Mingay, M., Blow, M. J., Zhang, Y., Yoshinaga, Y., Daum, C. G., & O'Malley, R. C. (2021). Persistence and plasticity in bacterial gene regulation. *Nature Methods*, *18*(12), 1499–1505. https://doi.org/10.1038/s41592-021-01312-2

Bentley, S. D., Chater, K. F., Cerdeño-Tárraga, A.-M., Challis, G. L., Thomson, N. R., James, K. D., Harris, D. E., Quail, M. A., Kieser, H., Harper, D., Bateman, A., Brown, S., Chandra, G., Chen, C. W., Collins, M., Cronin, A., Fraser, A., Goble, A., Hidalgo, J., … Hopwood, D. A. (2002). Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). *Nature*, *417*(6885), 141–147. https://doi.org/10.1038/417141a

Bertram, R., Rigali, S., Wood, N., Lulko, A. T., Kuipers, O. P., & Titgemeyer, F. (2011). Regulon of the N-Acetylglucosamine Utilization Regulator NagR in Bacillus subtilis. *Journal of Bacteriology*, *193*(14), 3525–3536. https://doi.org/10.1128/JB.00264-11

Book, A. J., Lewin, G. R., Mcdonald, B. R., Takasuka, T. E., Wendt-Pienkowski, E., Doering, D. T., Suh, S., Raffa, K. F., Fox, B. G., & Currie, C. R. (2016). Evolution of High Cellulolytic Activity in Symbiotic Streptomyces through Selection of Expanded Gene Content and Coordinated Gene Expression. *PLoS Biol*, *14*(6). https://doi.org/10.1371/journal.pbio.1002475

Browning, D. F., & Busby, S. J. W. (2016). Local and global regulation of transcription initiation in bacteria. *Nature Reviews Microbiology*, *14*(10), 638–650. https://doi.org/10.1038/nrmicro.2016.103

Browning, D. F., Butala, M., & Busby, S. J. W. (2019). Bacterial Transcription Factors: Regulation by Pick “N” Mix. *Journal of Molecular Biology*, *431*(20), 4067–4077. https://doi.org/10.1016/j.jmb.2019.04.011

Crooks, G. E., Hon, G., Chandonia, J.-M., & Brenner, S. E. (2004). WebLogo: A Sequence Logo Generator. *Genome Research*, *14*, 1188–1190.

Dwarakanath, S., Chaplin, A. K., Hough, M. A., Rigali, S., Vijgenboom, E., & Worrall, J. A. R. (2012). Response to Copper Stress in Streptomyces lividans Extends beyond Genes under Direct Control of a Copper-sensitive Operon Repressor Protein (CsoR) *. *Journal of Biological Chemistry*, *287*(21), 17833–17847. https://doi.org/10.1074/jbc.M112.352740

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLOS Computational Biology*, *7*(10), e1002195. https://doi.org/10.1371/journal.pcbi.1002195

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., Qureshi, M., Richardson, L. J., Salazar, G. A., Smart, A., Sonnhammer, E. L. L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S. C. E., & Finn, R. D. (2019). The Pfam protein families database in 2019. *Nucleic Acids Research*, *47*(D1), D427–D432. https://doi.org/10.1093/nar/gky995

Francis, I. M., Jourdan, S., Fanara, S., Loria, R., & Rigali, S. (2015). The cellobiose sensor CebR is the gatekeeper of Streptomyces scabies pathogenicity. *MBio*, *6*(2), e02018. https://doi.org/10.1128/mBio.02018-14

Giannotta, F., Georis, J., Moreau, A., Mazy-Servais, C., Joris, B., & Dusart, J. (1996). A sequence-specific DNA-binding protein interacts with the xZnC upstream region of Streptomyces sp. Strain EC3. *FEMS Microbiology Letters*, *142*, 91–97. https://doi.org/10.1111/j.1574-6968.1996.tb08413.x

Giannotta, F., Rigali, S., Virolle, M.-J., & Dusart, J. (2003). Site-directed mutagenesis of conserved inverted repeat sequences in the xylanase C promoter region from Streptomyces sp. EC3. *Molecular Genetics and Genomics*, *270*, 337–346. https://doi.org/10.1007/s00438-003-0927-y

Hodgson, D. A. (2000). Primary metabolism and its control in streptomycetes: A most unusual group of bacteria. In *Advances in Microbial Physiology* (Vol. 42, pp. 47–238). Academic Press. https://doi.org/10.1016/S0065-2911(00)42003-5

Ishihama, A., Shimada, T., & Yamazaki, Y. (2016). Transcription profile of Escherichia coli: Genomic SELEX search for regulatory targets of transcription factors. *Nucleic Acids Research*, *44*(5), 2058–2074. https://doi.org/10.1093/nar/gkw051

Janky, R., & van Helden, J. (2008). Evaluation of phylogenetic footprint discovery for predicting bacterial cis-regulatory elements and revealing their evolution. *BMC Bioinformatics*, *9*(1), 37. https://doi.org/10.1186/1471-2105-9-37

Jourdan, S., Francis, I. M., Kim, M. J., Salazar, J. J. C., Planckaert, S., Frère, J.-M., Matagne, A., Kerff, F., Devreese, B., Loria, R., & Rigali, S. (2016). The CebE/MsiK Transporter is a Doorway to the Cello-oligosaccharide-mediated Induction of Streptomyces scabies Pathogenicity. *Scientific Reports*, *6*(January), 27144. https://doi.org/10.1038/srep27144

Kazanov, M. D., Li, X., Gelfand, M. S., Osterman, A. L., & Rodionov, D. A. (2013). Functional diversification of ROK-family transcriptional regulators of sugar catabolism in the Thermotogae phylum. *Nucleic Acids Research*, *41*(2), 790–803. https://doi.org/10.1093/nar/gks1184

Lechner, M., Findeiß, S., Steiner, L., Marz, M., Stadler, P. F., & Prohaska, S. J. (2011). *Proteinortho: Detection of (Co-)orthologs in large-scale analysis*. https://doi.org/10.1186/1471-2105-12-124

Leyn, S. A., Suvorova, I. A., Kazakov, A. E., Ravcheev, D. A., Stepanova, V. V., Novichkov, P. S., & Rodionov, D. A. (2016). Comparative genomics and evolution of transcriptional regulons in Proteobacteria. *Microbial Genomics*, *2*(7). https://doi.org/10.1099/mgen.0.000061

Li, H., Rhodius, V., Gross, C., & Siggia, E. D. (2002). Identification of the binding sites of regulatory proteins in bacterial genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(18), 11772–11777. https://doi.org/10.1073/pnas.112341999

Liao, C., Rigali, S., Cassani, C. L., Marcellin, E., Nielsen, L. K., & Ye, B.-C. (2014). Control of chitin and N-acetylglucosamine utilization in Saccharopolyspora erythraea. *Microbiology*, *160*(9), 1914–1928. https://doi.org/10.1099/mic.0.078261-0

Liu, B., Yang, J., Li, Y., McDermaid, A., & Ma, Q. (2018). An algorithmic perspective of de novo cis-regulatory motif finding based on ChIP-seq data. *Briefings in Bioinformatics*, *19*(5), 1069–1081. https://doi.org/10.1093/bib/bbx026

Marushima, K., Ohnishi, Y., & Horinouchi, S. (2009). CebR as a Master Regulator for Cellulose/Cellooligosaccharide Catabolism Affects Morphological Development in

Streptomyces griseus. *Journal of Bacteriology*, *191*(19), 5930–5940. https://doi.org/10.1128/JB.00703-09

Mejía-Almonte, C., Busby, S. J. W., Wade, J. T., van Helden, J., Arkin, A. P., Stormo, G. D., Eilbeck, K., Palsson, B. O., Galagan, J. E., & Collado-Vides, J. (2020). Redefining fundamental concepts of transcription initiation in bacteria. *Nature Reviews Genetics*, *21*(11), 699–714. https://doi.org/10.1038/s41576-020-0254-8

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1), D412–D419. https://doi.org/10.1093/nar/gkaa913

Mwangi, M. M., & Siggia, E. D. (2003). Genome wide identification of regulatory motifs in Bacillus subtilis. *BMC Bioinformatics*, *4*, 18. https://doi.org/10.1186/1471-2105-4-18

Nguyen, J. (1999). The regulatory protein Reg1 of *Streptomyces lividans* binds the promoter region of several genes repressed by glucose. *FEMS Microbiology Letters*, *175*(1), 51–58. https://doi.org/10.1111/j.1574-6968.1999.tb13601.x

Nguyen, J., Francou, F., Virolle, M. J., & Guérineau, M. (1997). Amylase and chitinase genes in Streptomyces lividans are regulated by reg1, a pleiotropic regulatory gene. *Journal of Bacteriology*, *179*(20), 6383–6390. https://doi.org/10.1128/jb.179.20.6383-6390.1997

Novichkov, P. S., Kazakov, A. E., Ravcheev, D. A., Leyn, S. A., Kovaleva, G. Y., Sutormin, R. A., Kazanov, M. D., Riehl, W., Arkin, A. P., Dubchak, I., & Rodionov, D. A. (2013). RegPrecise 3.0—A resource for genome-scale exploration of transcriptional regulation in bacteria. *BMC Genomics*, *14*(1), 1. https://doi.org/10.1186/1471-2164-14-745

Novichkov, P. S., Laikova, O. N., Novichkova, E. S., Gelfand, M. S., Arkin, A. P., Dubchak, I., & Rodionov, D. A. (2010). RegPrecise: A database of curated genomic inferences of transcriptional regulatory interactions in prokaryotes. *Nucleic Acids Research*, *38*(Database issue), D111-8. https://doi.org/10.1093/nar/gkp894

Ohashi, K., Hataya, S., Nakata, A., Matsumoto, K., Kato, N., Sato, W., Carlos-Shanley, C., Beebe, E. T., Currie, C. R., Fox, B. G., & Takasuka, T. E. (2021). Mannose- and Mannobiose-Specific Responses of the Insect-Associated Cellulolytic Bacterium Streptomyces sp. Strain SirexAA-

E. *Applied and Environmental Microbiology*, *87*(14), e02719-20. https://doi.org/10.1128/AEM.02719-20

Ortet, P., De Luca, G., Whitworth, D. E., & Barakat, M. (2012). P2TF: a comprehensive resource for analysis of prokaryotic transcription factors. *BMC Microbiology*, *13*(628). http://www.p2tf.org/.

Park, P. J. (2009). ChIP–seq: Advantages and challenges of a maturing technology. *Nature Reviews Genetics*, *10*(10), 669–680. https://doi.org/10.1038/nrg2641

Ravcheev, D. A., Khoroshkin, M. S., Laikova, O. N., Tsoy, O. V., Sernova, N. V., Petrova, S. A., Rakhmaninova, A. B., Novichkov, P. S., Gelfand, M. S., & Rodionov, D. A. (2014). Comparative genomics and evolution of regulons of the LacI-family transcription factors. *Frontiers in Microbiology*, *5*, 294. https://doi.org/10.3389/fmicb.2014.00294

Rigali, S., Derouaux, A., Giannotta, F., & Dusart, J. (2002). Subdivision of the helix-turn-helix GntR family of bacterial regulators in the FadR, HutC, MocR, and YtrA subfamilies. *Journal of Biological Chemistry*, *277*(15), 12507–12515. https://doi.org/10.1074/jbc.M110968200

Rigali, S., Schlicht, M., Hoskisson, P., Nothaft, H., Merzbacher, M., Joris, B., & Titgemeyer, F. (2004). 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 Research*, *32*(11), 3418–3426. https://doi.org/10.1093/nar/gkh673

Rodionov, D. A. (2007). Comparative Genomic Reconstruction of Transcriptional Regulatory Networks in Bacteria. *Chemical Reviews*, *107*(8). https://doi.org/10.1021/cr068309

Santos-Zavaleta, A., Salgado, H., Gama-Castro, S., Sánchez-Pérez, M., Gómez-Romero, L., Ledezma-Tejeida, D., García-Sotelo, J. S., Alquicira-Hernández, K., Muñiz-Rascado, L. J., Peña-Loredo, P., Ishida-Gutiérrez, C., Velázquez-Ramírez, D. A., Del Moral-Chávez, V., Bonavides-Martínez, C., Méndez-Cruz, C.-F., Galagan, J., & Collado-Vides, J. (2019). RegulonDB v 10.5: Tackling challenges to unify classic and high throughput knowledge of gene regulation in *E. coli* K-12. *Nucleic Acids Research*, *47*(D1), D212–D220. https://doi.org/10.1093/nar/gky1077

Schlösser, A., Aldekamp, T., Schrempf, H., R.G., B., H.C., P., M.A., S., R.G., B., & P., L. (2000). Binding characteristics of CebR, the regulator of the *ceb* operon required for cellobiose/cellotriose uptake in *Streptomyces reticuli*. *FEMS Microbiology Letters*, *190*(1), 127–132. https://doi.org/10.1111/j.1574-6968.2000.tb09274.x

Schlösser, A., Weber, A., Schrempf, H., M, K., B, B., W, B., R.G., B., & P, L. (2001). Synthesis of the *Streptomyces lividans* maltodextrin ABC transporter depends on the presence of the regulator MalR. *FEMS Microbiology Letters*, *196*(1), 77–83. https://doi.org/10.1111/j.1574-6968.2001.tb10544.x

Studholme, D. J., Bentley, S. D., & Kormanec, J. (2004). Bioinformatic identification of novel regulatory DNA sequence motifs in Streptomyces coelicolor. *BMC Microbiology*, *4*(1), 14. https://doi.org/10.1186/1471-2180-4-14

Świątek-Połatyńska, M. A., Bucca, G., Laing, E., Gubbens, J., Titgemeyer, F., Smith, C. P., Rigali, S., & Wezel, G. P. van. (2015). Genome-Wide Analysis of In Vivo Binding of the Master Regulator DasR in Streptomyces coelicolor Identifies Novel Non-Canonical Targets. *PLOS ONE*, *10*(4), e0122479. https://doi.org/10.1371/journal.pone.0122479

Tenconi, E., Urem, M., Świątek-Połatyńska, M. A., Titgemeyer, F., Muller, Y. A., van Wezel, G. P., & Rigali, S. (2015). Multiple allosteric effectors control the affinity of DasR for its target sites. *Biochemical and Biophysical Research Communications*, *464*(1), 324–329. https://doi.org/10.1016/j.bbrc.2015.06.152

Tsevelkhoroloo, M., Shim, S. H., Lee, C.-R., Hong, S.-K., & Hong, Y.-S. (2021). LacI-Family Transcriptional Regulator DagR Acts as a Repressor of the Agarolytic Pathway Genes in Streptomyces coelicolor A3(2). *Frontiers in Microbiology*, *12*, 658657. https://doi.org/10.3389/fmicb.2021.658657

Tsujibo, H., Kosaka, M., Ikenishi, S., Sato, T., Miyamoto, K., & Inamori, Y. (2004). Molecular characterization of a high-affinity xylobiose transporter of Streptomyces thermoviolaceus OPC-520 and its transcriptional regulation. *Journal of Bacteriology*, *186*(4), 1029–1037. https://doi.org/10.1128/JB.186.4.1029-1037.2004

Urem, M., van Rossum, T., Bucca, G., Moolenaar, G. F., Laing, E., Świątek-Połatyńska, M. A., Willemse, J., Tenconi, E., Rigali, S., Goosen, N., Smith, C. P., & van Wezel, G. P. (2016). OsdR of *Streptomyces coelicolor* and the Dormancy Regulator DevR of *Mycobacterium tuberculosis* Control Overlapping Regulons. *MSystems*, *1*(3), e00014-16. https://doi.org/10.1128/mSystems.00014-16

van der Meij, A., Worsley, S. F., Hutchings, M. I., & van Wezel, G. P. (2017). Chemical ecology of antibiotic production by actinomycetes. *FEMS Microbiology Reviews*, *41*(3), 392–416. https://doi.org/10.1093/femsre/fux005

Van Hijum, S. A. F. T., Medema, M. H., & Kuipers, O. P. (2009). Mechanisms and Evolution of Control Logic in Prokaryotic Transcriptional Regulation. *MICROBIOLOGY AND MOLECULAR BIOLOGY REVIEWS*, *73*(3), 481–509. https://doi.org/10.1128/MMBR.00037-08

van Wezel, G. P., White, J., Bibb, M. J., & Postma, P. W. (1997). The malEFG gene cluster of Streptomyces coelicolor A3(2): Characterization, disruption and transcriptional analysis. *Molecular and General Genetics MGG*, *254*(5), 604–608. https://doi.org/10.1007/s004380050458

van Wezel, G. P., White, J., Young, P., Postma, P. W., & Bibb, M. J. (1997). 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. *Molecular Microbiology*, *23*(3), 537–549. https://doi.org/10.1046/j.1365-2958.1997.d01-1878.x

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, *10*(1), 57–63. https://doi.org/10.1038/nrg2484

Wasserman, W. W., & Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, *5*(4), 276–287. https://doi.org/10.1038/nrg1315

Yao, L.-L., Liao, C.-H., Huang, G., Zhou, Y., Rigali, S., Zhang, B., & Ye, B.-C. (2014). GlnR-mediated regulation of nitrogen metabolism in the actinomycete Saccharopolyspora erythraea. *Applied Microbiology and Biotechnology*, *98*(18), 7935–7948. https://doi.org/10.1007/s00253-014-5878-1

SUPPLEMENTARY DATA

Supplementary Figure S1

Figure S1. Number of LacI TFs per *Streptomyces* species. Note the extreme cases of *S. bingchenggensis* (69 LacI TFs) and *S. olivoreticuli* (6 LacI TFs)

Supplementary Table S1

Table S1. List of plasmids used in this study					
COG	**Representative Gene**	**Protein_ID**	**Plasmid**	**Insertion**	**Vector**
LacI001	SCO3943	NP_628127.1	pSIN027	NdeI_XhoI	pET-28a(+)
LacI002	SCO4158	NP_628335.1	pSIN028	NdeI_XhoI	pET-28a(+)
LacI003	SCO1078	NP_625372.1	pSIN002	NdeI_HindIII	pET-22b
LacI007	SCO6713	NP_630786.1	pSIN033	NdeI_XhoI	pET-28a(+)
LacI008	SCO2753	NP_626984.1	pSIN026	NdeI_XhoI	pET-28a(+)
LacI009	SCO0886	NP_625185.1	pSIN020	NdeI_XhoI	pET-28a(+)
LacI010	SCO2745	NP_626976.1	pSIN025	NdeI_XhoI	pET-28a(+)
LacI011	SCO7014	NP_631078.1	pSIN034	NdeI_XhoI	pET-28a(+)
LacI014	SCO5692	NP_629820.1	pSIN030	NdeI_XhoI	pET-28a(+)
LacI015	SCO7027	NP_631091.1	pSIN035	NdeI_XhoI	pET-28a(+)
LacI017	SCO6598	NP_630677.1	pSIN032	NdeI_XhoI	pET-28a(+)
LacI018	SCO1956	NP_626220.1	pSIN024	NdeI_XhoI	pET-28a(+)
LacI019	SCO1376	NP_625660.1	pSIN029	NdeI_XhoI	pET-28a(+)
LacI023	SBI_RS45370	WP_014181788.1	pSIN007	NdeI_XhoI	pET-28a(+)
LacI024	SCO7411	NP_631460.1	pSIN036	NdeI_XhoI	pET-28a(+)
LacI025	SCO7502	NP_631548.1	pSIN037	NdeI_XhoI	pET-28a(+)
LacI026	SBI_RS08050	WP_043486265.1	pSIN005	NdeI_XhoI	pET-28a(+)
LacI027	SCO1066	NP_625360.1	pSIN004	NdeI_XhoI	pET-28a(+)
LacI028	SCO7554	NP_631597.1	pSIN038	NdeI_XhoI	pET-28a(+)
LacI029	SCO6233	NP_630334.1	pSIN031	NdeI_XhoI	pET-28a(+)
LacI031	SGR_RS11925	WP_012379184.1	pSIN040	NdeI_XhoI	pET-28a(+)
LacI032	SCO0629	NP_624940.1	pSIN019	NdeI_XhoI	pET-28a(+)
LacI033	SBI_RS48885	WP_043492894.1	pSIN010	NdeI_XhoI	pET-28a(+)
LacI035	SCO0289	NP_624618.1	pSIN017	NdeI_XhoI	pET-28a(+)
LacI036	SBI_RS46210	WP_014181967.1	pSIN039	NdeI_XhoI	pET-28a(+)
LacI037	SCO0456	NP_624776.1	pSIN018	NdeI_XhoI	pET-28a(+)
LacI038	SCAB_RS26610	WP_013003177.1	pSIN014	NdeI_XhoI	pET-28a(+)
LacI039	SGR_RS17280	WP_012379935.1	pSIN044	NdeI_XhoI	pET-28a(+)
LacI042	SCAB_RS02460	WP_012998467.1	pSIN006	NdeI_XhoI	pET-28a(+)
LacI044	SBI_RS01965	WP_014173018.1	pSIN008	NdeI_XhoI	pET-28a(+)
LacI050	WQO_RS32820	WP_029182469.1	pSIN045	NdeI_XhoI	pET-28a(+)
LacI051	SBI_RS21665	WP_014177014.1	pSIN051	NdeI_XhoI	pET-28a(+)
LacI053	SGR_RS04505	WP_012378184.1	pSIN043	NdeI_XhoI	pET-28a(+)
LacI059	SBI_RS36130	WP_043487665.1	pSIN013	NdeI_XhoI	pET-28a(+)
LacI064	SBI_RS03795	WP_014173393.1	pSIN046	NdeI_XhoI	pET-28a(+)
LacI066	SBI_RS46800	WP_043492711.1	pSIN049	NdeI_XhoI	pET-28a(+)
LacI071	SCAB_RS41390	WP_037728820.1	pSIN012	NdeI_XhoI	pET-28a(+)
LacI072	STRVI_RS24580	WP_106685713.1	pSIN047	NdeI_XhoI	pET-28a(+)
LacI073	SVTN_RS35145	WP_041132710.1	pSIN052	NdeI_XhoI	pET-28a(+)
LacI077	SBI_RS06345	WP_043489044.1	pSIN041	NdeI_XhoI	pET-28a(+)
LacI102	SVTN_RS01870	WP_052498857.1	pSIN003	NdeI_XhoI	pET-28a(+)

Supplementary Table S2

Table S2. List of primers used in this study

Primer name	LacI COG	Sequence (5'-3')
ManR_emsa_F	003	TCACCGCTTT**GACAACGTTGTC**AGATACCGCA
ManR_emsa_R		TGCGGTATCT**GACAACGTTGTC**AAAGCGGTGATGGTACACGCACCCTGTCGT
1_emsa_F	102	TCACCGCTTT**TGCCAAAACGTTTTCGCA**AGATACCGCA
1_emsa_R		TGCGGTATCT**TGCGAAAACGTTTTGGCA**AAAGCGGTGATGGTACACGCACCCTGTCGT
2_emsa_F	027	TCACCGCTTT**TTACAACGTTGTAA**AGATACCGCA
2_emsa_R		TGCGGTATCT**TTACAACGTTGTAA**AAAGCGGTGATGGTACACGCACCCTGTCGT
5_emsa_F	023	TCACCGCTTT**TGGAcCGGTCCA**AGATACCGCA
5_emsa_R		TGCGGTATCT**TGGACCGgTCCA**AAAGCGGTGATGGTACACGCACCCTGTCGT
6_emsa_F	044	TCACCGCTTT**TCGAACCGGTTCGa**AGATACCGCA
6_emsa_R		TGCGGTATCT**tCGAACCGGTTCGA**AAAGCGGTGATGGTACACGCACCCTGTCGT
10_emsa_F	071	TCACCGCTTT**caCTAATACGTaTTAGtT**AGATACCGCA
10_emsa_R		TGCGGTATCT**AaCTAAtACGTATTAGtg**AAAGCGGTGATGGTACACGCACCCTGTCGT
11_emsa_F	059	TCACCGCTTT**TGTGACCGGTCACA**AGATACCGCA
11_emsa_R		TGCGGTATCT**TGTGACCGGTCACA**AAAGCGGTGATGGTACACGCACCCTGTCGT
16_emsa_F	037	TCACCGCTTT**tTTAAACCGGTTtAAA**AGATACCGCA
16_emsa_R		TGCGGTATCT**TTTaAACCGGTTTAAa**AAAGCGGTGATGGTACACGCACCCTGTCGT
19_emsa_F	027	TCACCGCTTT**TTACAACGTTGTAA**AGATACCGCA
19_emsa_R		TGCGGTATCT**TTACAACGTTGTAA**AAAGCGGTGATGGTACACGCACCCTGTCGT
21_emsa_F	019	TCACCGCTTT**CGAATGTTCCGG**AGATACCGCA
21_emsa_R		TGCGGTATCT**CCGGAACATTCG**AAAGCGGTGATGGTACACGCACCCTGTCGT
23_emsa_F	010	TCACCGCTTT**TGTAATCGATTCCA**AGATACCGCA
23_emsa_R		TGCGGTATCT**TGGAATCGATTACA**AAAGCGGTGATGGTACACGCACCCTGTCGT
24_emsa_F	008	TCACCGCTTT**AGcAAGCGCTTTCT**AGATACCGCA
24_emsa_R		TGCGGTATCT**AGAAAGCGCTTgCT**AAAGCGGTGATGGTACACGCACCCTGTCGT
25_emsa_F	001	TCACCGCTTT**GAGCCCTACTATCGGCTC**AGATACCGCA
25_emsa_R		TGCGGTATCT**GAGCCGATAGTAGGGCTC**AAAGCGGTGATGGTACACGCACCCTGTCGT
39_emsa_F	077	TCACCGCTTT**ATGTTGCAACGTTGCAAGCA**AGATACCGCA
39_emsa_R		TGCGGTATCT**TGCTTGCAACGTTGCAACAT**AAAGCGGTGATGGTACACGCACCCTGTCGT
43_emsa_F	050	TCACCGCTTT**TgCAACGTTGcA**AGATACCGCA
43_emsa_R		TGCGGTATCT**TgCAACGTTGcA**AAAGCGGTGATGGTACACGCACCCTGTCGT
45_emsa_F	072	TCACCGCTTT**CTAGAACGTTcTAG**AGATACCGCA
45_emsa_R		TGCGGTATCT**CTAgAACGTTCTAG**AAAGCGGTGATGGTACACGCACCCTGTCGT
50_emsa_F	073	TCACCGCTTT**TGCGGGAACGTTCCCGCA**AGATACCGCA
50_emsa_R		TGCGGTATCT**TGCGGGAACGTTCCCGCA**AAAGCGGTGATGGTACACGCACCCTGTCGT
Random_Cy5	-	ACGACAGGGTGCGTGTACCA