Published online June 24, 2004

3418–3426 Nucleic Acids Research, 2004, Vol. 32, No. 11
DOI: 10.1093/nar/gkh673

Extending the classification of bacterial transcription factors beyond the helix–turn–helix motif as an alternative approach to discover new *cis*/*trans* relationships

Sébastien Rigali*, Maximilian Schlicht[1], Paul Hoskisson[2], Harald Nothaft[1], Matthias Merzbacher[1], Bernard Joris and Fritz Titgemeyer[1]

Centre d'Ingénierie des Protéines, Université de Liège, Institut de Chimie B6a, B-4000, Liège, Belgium, [1]Lehrstuhl für Mikrobiologie, Friedrich-Alexander-Universität Erlangen-Nürnberg, Staudtstrasse 5, 91058 Erlangen, Germany and [2]Department of Molecular Microbiology, John Innes Centre, Colney, Norwich NR4 7UH, UK

Received May 5, 2004; Revised and Accepted June 8, 2004

ABSTRACT

Transcription factors (TFs) of bacterial helix–turn–helix superfamilies exhibit different effector-binding domains (EBDs) fused to a DNA-binding domain with a common feature. In a previous study of the GntR superfamily, we demonstrated that classifying members into subfamilies according to the EBD heterogeneity highlighted unsuspected and accurate TF-binding site signatures. In this work, we present how such *in silico* analysis can provide prediction tools to discover new *cis*/*trans* relationships. The TF-binding site consensus of the HutC/GntR subfamily was used to (i) predict target sites within the *Streptomyces coelicolor* genome, (ii) discover a new HutC/GntR regulon and (iii) discover its specific TF. By scanning the *S.coelicolor* genome we identified a presumed new HutC regulon that comprises genes of the phosphotransferase system (PTS) specific for the uptake of *N*-acetylglucosamine (PTS^{Nag}). A weight matrix was derived from the compilation of the predicted *cis*-acting elements upstream of each gene of the presumed regulon. Under the assumption that TFs are often subject to autoregulation, we used this matrix to scan the upstream region of the 24 HutC-like members of *S.coelicolor*. *orf* SCO5231 (*dasR*) was selected as the best candidate according to the high score of a 16 bp sequence identified in its upstream region. Our prediction that DasR regulates the PTS^{Nag} regulon was confirmed by *in vivo* and *in vitro* experiments. In conclusion, our *in silico* approach permitted to highlight the specific TF of a regulon out of the 673 *orfs* annotated as 'regulatory proteins' within the genome of *S.coelicolor*.

INTRODUCTION

In the prokaryotic world, the most-studied and best-characterized group of transcription factors (TFs) are those that contain the helix–turn–helix (HTH) DNA-binding motif (1). Protein families of the HTH group have been identified mainly through sequence comparisons focused almost exclusively on the HTH structure within the DNA-binding domain (DBD) (2–4). This classification strategy is entirely justified, as the HTH motif is the only region that shows strong similarities among all members of the group and thus provides the basis for a simple method of classification and detection of new members.

Only a small number of studies have attempted to identify specific signatures beyond the HTH motif in the two other components involved in the transcriptional process. These are the effector-binding domain (EBD) and the *cis*-acting elements (5–10). Pattern search in the three components is particularly difficult when members of a superfamily exhibit important EBD heterogeneity fused to a common DBD, as is the case, e.g., for TFs of the LysR (11) or GntR superfamilies (9,10). This problem can be described from a structural point of view as one could imagine that different EBDs impose different sterical constraints on a common feature of the DBD, which in turn may impose various orientations and presentations of the HTH motif, ultimately reflected by the accommodation of *cis*-acting elements. Hence, with superfamilies, the compilation of the recognized DNA sequences can result in an inaccurate consensus signature, which is inappropriate for use in predicting putative *cis*-acting elements in a given genome. A good example of this is the consensus sequence for the *cis*-acting elements recognized by the GntR superfamily's members, GT-N(0–15)-AC (10), which cannot be used as an operator site prediction tool as the sequence is found in the upstream regions of almost all genes in bacteria. However, we showed that this problem could be improved by defining the limits of the EBD multiplicity through a combination of similarity analysis and

*To whom correspondence should be addressed. Tel: +32 4 366 33 77; Fax: +32 4 366 33 64; Email: srigali@ulg.ac.be

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors secondary structure topology prediction (10). We have investigated this method with the GntR superfamily, one of the most represented families in the prokaryotic world, that currently comprises more than 1300 sequenced members (February 2004). We limited the C-terminal EBD domain heterogeneity to four major subfamilies that we called FadR, HutC, MocR and YtrA, which make 95% of all members. The remaining members belong to other minor subfamilies such as AraR (10,12), PlmA (13), or those that are not yet characterized. Once members were separated into subfamilies, we were able to suggest new *cis*-element consensus sites for each subfamily. That is how the *cis*-element consensi of the FadR subfamily (T.GT-$N_{(0-3)}$-AC.T) and HutC subfamily (GT-N(1)-TA-N(1)-AC) were highlighted and proposed as binding site prediction tools of the respective TF subfamilies (10).

Predicting gene transcription regulation is currently one of the great challenges of molecular biology as it provides a complementary analysis to genomic approaches such as transcriptome and proteome studies to the understanding of new regulatory systems. As a result, databases of transcription factors and their genomic binding sites in *Escherichia coli*, such as DPInteract or RegulonDB, have been created (14,15) and several computer algorithms for representation and discovery of DNA-binding sites have been reported (16). To understand and predict regulatory systems within a bacterial genome, the identification of the regulatory proteins is considered as the first and easiest task as there are good methods available to group proteins into families (17). However, through our analysis of the GntR superfamily, we demonstrated the limit of the actual classification of TFs as it reduces and falsifies the information of the deduced TF-binding site consensus or the weight matrices. If one could extend the classification of HTH superfamilies beyond the HTH motif, it should be feasible to unravel the true relationships between *cis*- and *trans*-acting elements by minimizing the risk to compile sequences that should not be included together in a multiple alignment.

In this paper, we present such an *in silico* approach to derive a more accurate prediction of new regulatory codes from genomic information, which we have validated experimentally. We chose the theoretical data and prediction tools available for the HutC/GntR subfamily (10,18) and the *Streptomyces coelicolor* genome as a model for the experimental validation of the predicted *cis*/*trans* regulatory code.

MATERIALS AND METHODS

Selection of members of the GntR superfamily and classification into subfamilies

GntR-like TFs are classified according to the PROSITE weight matrix of the HTH motif: entry PS50949. GntR members from *S.coelicolor* genome [(19); accession no.: NC_003888] were selected by keywords using the Sequence Retrieval System (SRS) available at the ExPASy (Expert Protein Analysis System) Molecular Biology Server: http://www.expasy.org/, Expasy home page. The classification of selected members into the various GntR subfamilies was obtained from secondary structure predictions as described previously (10).

Prediction of the HutC-like transcription factor binding sites in *S.coelicolor* genome

The HutC subfamily DNA consensus pattern GT-N(1)-TA-N(1)-AC (10) was used as prediction tool to search the *S.coelicolor* genomes for this DNA signature in the intergenic regions. Therefore, we used the PATTERN SEARCH program available at the *S.coelicolor* genome server site: http://jiio16.jic.bbsrc.ac.uk/S.coelicolor/, home page. Functionally related *orfs* were searched to find a set of coregulated genes. *orfs* SCO1390 (*crr*), SCO2905c (*malX2*), SCO2907 (*nagE2*) and SCO5841c (*ptsH*) were selected. They encode enzyme IIA^{Crr}, enzyme IIB^{Nag}, enzyme IIC^{Nag} and HPr of the *N*-acetylglucosamine-specific PTS (20,21).

Construction of alignment and weight matrices of PTS co-regulated genes

We chose the *cis*-elements upstream of *orfs* SCO1390, SCO2905c, SCO2907, and SCO5841c to build alignment and weight matrices specific for PTS^{Nag} regulon by using the Target Explorer automated tool (22) available at http://trantor.bioc.columbia.edu/Target_Explorer/, Target Explorer home page. The alignment matrix lists the number of occurrences of each letter at each position of an alignment while in the weight matrix, the elements are the weights used to score a test sequence to measure how close that sequence word matches the pattern described by the matrix. A test sequence is compared to the weight matrix, and its score is the sum of the weights for the letter aligned at each position. The alignment matrix was translated into a weight matrix using the expression

$$\text{weight}_{ij} = \ln\left\{\frac{\left[\left(n_{ij} + pi\right)/(N+1)\right]}{pi}\right\} \sim \ln\left(\frac{f_{ij}}{pi}\right)$$

where N is the total number of sequences in the alignment, n_{ij} is the number of times nucleotide i is observed in position j of the alignment, $f_{ij} = n_{ij} /N$ is the frequency of letter i at position j, pi is the *a priori* probability of letter i in the sets of the input DNA sequences. A positive weight_{ij} implies that the frequency of letter i at position j of the alignment is higher than the *a priori* probability of this letter. The matrix specific for PTS^{Nag} regulon has been saved in the public library under the 'PTScoeli' denomination.

Search for best HutC-like candidates involved in PTS genes regulation

To determine the best HutC-like candidate presumed to be involved in PTS regulation, we tested the DNA sequence upstream the 24 HutC-like regulators against the 'PTScoeli' weight matrix. By this way, we searched to highlight DNA patterns that suggest autoregulation of the presumed PTS^{Nag} regulon transcription factor. The sequence of the upstream region of each of the HutC-like regulators was selected using programs from the National Center for Biotechnology Information (NCBI). Once the 'PTScoeli' weight matrix was built, we fixed a low cut-off score (4.00 bits) to allow the prediction of *cis*-elements with mismatches in one or more conserved positions of the HutC/GntR consensus. The program PATSER (available at http://trantor.bioc.columbia.edu/Target_Explorer/, Target Explorer home page) was implemented to score individual potential PTS-like binding sites against the matrix (23). The predicted HutC-like transcription factor whose upstream region presented the best score against the weight matrix was then selected for experimental confirmation.

Construction of *dasR* expression plasmids

For overproduction of *dasR* in *E.coli*, *dasR* was amplified by PCR using *S.coelicolor* M145 wild-type chromosomal DNA as template together with oligonucleotides engineered to introduce the restriction sites NdeI and BamHI, respectively (MS3, ATATATACATATGAGCACCGACGTCAGCAGTGC and MS4, AAGGATCCGTGATGATGATGATGATGGTCCTGGGGCCGCTTGAGG; restriction sites underlined). Total DNA from *S.coelicolor* M145 was isolated as described (24). The oligonucleotide MS4 was designed with codons for a C-terminal hexahistidine (His)-tag. The product of the PCR was cut with NdeI and BamHI and ligated into the NdeI- and BamHI-digested plasmid pET3c (25), resulting in pFT240. The sequence of the inserted fragment was verified by DNA sequencing.

For overproduction of *dasR* in *S.coelicolor*, the cloning was achieved by PCR with oligonucleotides MS1 and MS2 (GCTTAATTAACTGAAAGGAGGTTAATAATGAGCACCGACGTCAGCAGTGC-3′; AGGGATCCTAGTCCTGGGGCCGCTTGAGGCGGGCCACG-3′, restriction sites are underlined). The PCR product was digested with PacI and BamHI and subcloned into the high-copy shuttle-vector derivative pUWL-SK+ pFT74 (K. Mahr, unpublished data), in which the *dasR* gene was placed under control of the constitutive glucose kinase gene promoter (P*glkA*). The resulting *dasR* expression plasmid was confirmed by DNA sequencing and then designated pFT241 (*dasR*$^+$). pFT241 was transferred into *S.coelicolor* wild-type M145 by protoplast transformation.

Production and purification of DasR

For the purification of recombinant His-tagged DasR, *E.coli* Rosetta (DE3) was transformed with pFT240 (Novagen, UK). Cells were grown at 37°C until the culture reached an OD_{600} of 0.35. Gene expression was induced by a final concentration of 1 mM isopropyl thio-β-D-galactoside. Incubation was continued for 2–3 h. Cells were harvested by centrifugation, washed, and ruptured by sonication in 20 mM Na_2HPO_4 (pH 7.2), 0.5 mM NaCl (start buffer). A fraction of soluble proteins was obtained after centrifugation at 4°C for 30 min at 14,000 *g* and was loaded onto a Ni^{2+}-NTA-agarose column (3 ml bed volume). The major fraction of His-tagged DasR protein eluted between 250 and 350 mM imidazole with a final yield of 0.5 mg homogeneous pure protein from 1 l bacterial culture. His-tagged DasR was dialysed against start buffer and stored at 4°C for further use.

Electrophoretic mobility shift assay

Purified His-tagged DasR protein (1 μg) was incubated at 30°C for 15 min with 25 pmol of DNA representing a potential *cis*-element and 1 μg non-specific DNA (PCR-script cloning vector) in a total volume of 20 μl containing 50 mM Tris–HCl, pH 7.5, 20 mM NaCl, 1 mM EDTA and 1 mM DTT. Reaction mixtures were supplemented with 5 μl glycerol before loading on a 1% (w/v) agarose gel. Bound and unbound probes were separated by conventional gel electrophoresis at room temperature and DNA was visualized by ethidium bromide staining. The predicted potential *cis*-acting elements were taken from the promoter regions of *crr* (SCO1390; 36 bp; 5′-CCGTGAGGAGTGTGGTCTAGACCTCTAATCGGAAC A-3′; probe I), *malX2* (SCO2905c; 48 bp; 5′-ACGGCGGTGCCGTCTGTCAACTGGTCTACACCAGTGTACCGGCGACCG-3′; probe II), *nagE2* (SCO2907; 37 bp; 5′-CAACAGGTCTACACCACTGAGTGGTGTAGACCACCAC-3′; probe III), *ptsH* (SCO5841c; 33 bp; 5′-AAACTGGTCTAGACAAGACTGGTCTAGACAACT-3′; probe IV), and *dasR* (SCO5231c; 49 bp; 5′-AGTCCTATTGGTCTGGGCCAAGCTCCCCGTACTGGTCTACACCATTGGT-3′; probe VI) respectively (nucleotides that constitute the HutC/GntR-binding site consensus sequence are underlined). The putative *crp* *cis*-acting element was applied as a non-specific control (SH1, 5′-TGCGGCATCCTTGTGACAGATCACACTGTTTGGACT-3′; probe V) (26).

N-acetylglucosamine uptake assay

Uptake of *N*-[^{14}C]acetyl-D-glucosamine (6.2 mCi mmol^{-1}) at a final concentration of 20 μM into mycelia was performed as described previously (20).

RESULTS

Overall strategy of the *cis*/*trans* prediction approach

Figure 1 shows a schematic presentation of the different steps (1 to 13) of our prediction approach which yields two goals: (i) prediction of target sites and identification if new regulons controlled by HutC/GntR members and (ii) selection of the best HutC/GntR candidate for a given gene, operon, or regulon followed by experimental confirmation. The procedure begins with an initial bioinformatics analysis to collect all members of a superfamily followed by classification into subfamilies (steps 1 to 3). A consensus sequence of TF-binding sites, which is here (GT-N(1)-TA-N(1)-AC) for the HutC subfamily, may be derived for some subfamilies (step 4). All members will then be identified in the genome of interest (step 5). The prediction tool PATTERN SEARCH will be applied to obtain a list of all putative HutC-like binding sites in a given genome (step 6). The presumed function of the genes downstream each predicted *cis* site will be annotated (step 7) and the set of genes involved in a common biological process are identified (steps 8 and 9). At this stage, the information is generated that allows the selection of the best TF for a given gene, operon or regulon. That is the number of members of the subfamily (24 HutC-like candidates for *S.coelicolor*). The accuracy of candidate prediction can be further narrowed by making two assumptions: (i) the regulatory gene is located in the vicinity of the target gene(s) and/or (ii) the regulatory gene is autoregulated. If for the latter the searched TF is for a single gene or operon, its predicted HutC-like *cis*-element will be aligned with the upstream regions of all HutC-like member genes to identify similar *cis* sequences (step 12). If the searched TF is responsible for a regulon, a weight matrix based on the compilation of all predicted target sites can be built by the CONSENSUS program (step 10). Similar *cis* patterns are scanned in the collected upstream regions of all HutC candidates by the PATSER software (step 12), which ultimately reveals a list

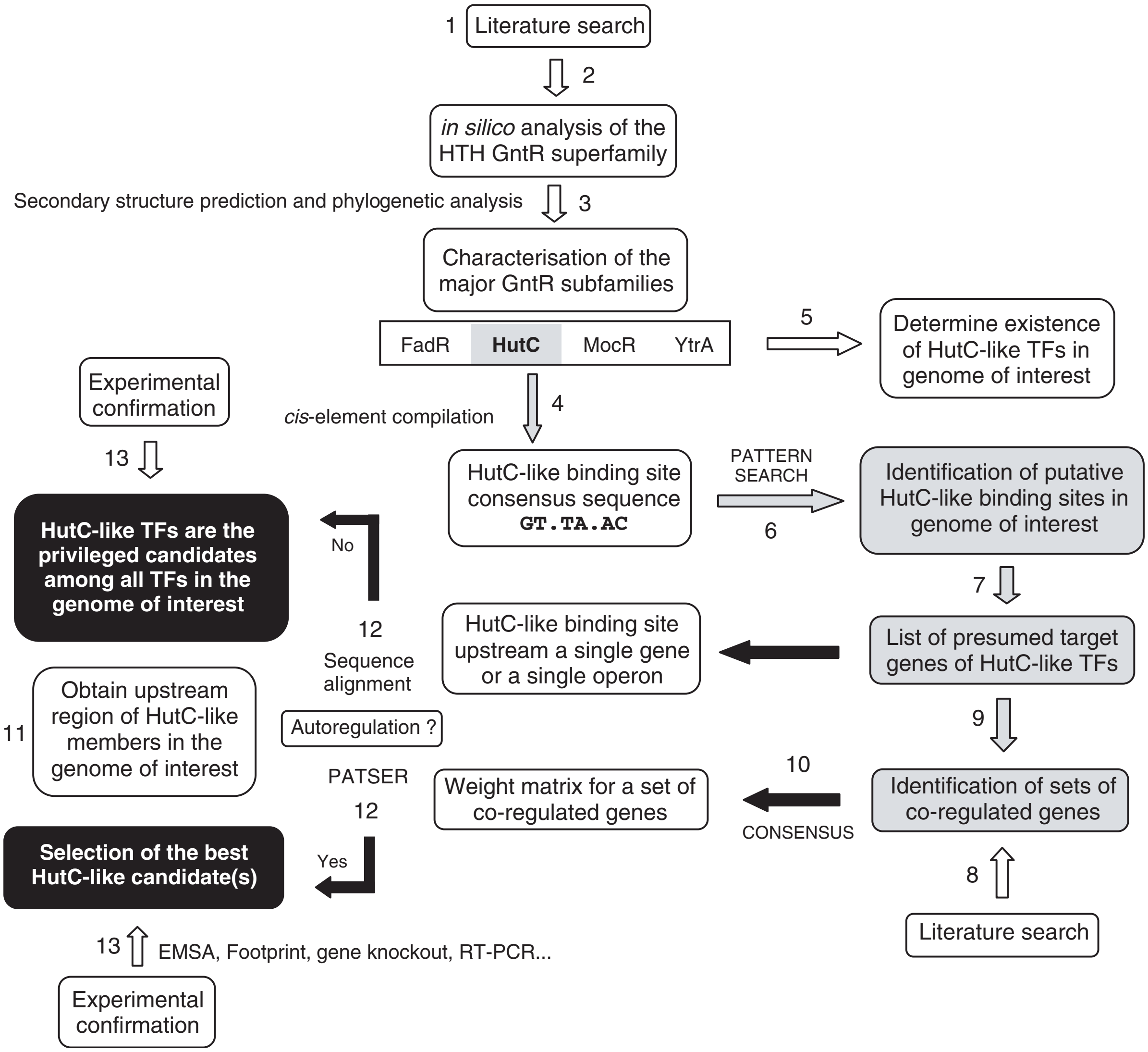
1 Literature search
2
in silico analysis of the HTH GntR superfamily
Secondary structure prediction and phylogenetic analysis
3
Characterisation of the major GntR subfamilies
FadR
HutC
MocR
YtrA
5
Determine existence of HutC-like TFs in genome of interest
Experimental confirmation
cis-element compilation
4
13
HutC-like binding site consensus sequence GT.TA.AC
PATTERN SEARCH
6
Identification of putative HutC-like binding sites in genome of interest
HutC-like TFs are the privileged candidates among all TFs in the genome of interest
No
7
12
HutC-like binding site upstream a single gene or a single operon
List of presumed target genes of HutC-like TFs
Sequence alignment
Obtain upstream region of HutC-like members in the genome of interest
11
Autoregulation ?
9
10
PATSER
Weight matrix for a set of co-regulated genes
Identification of sets of co-regulated genes
12
CONSENSUS
Selection of the best HutC-like candidate(s)
Yes
8
Literature search
13
EMSA, Footprint, gene knockout, RT-PCR...
Experimental confirmation

Figure 1. Overall strategy for the prediction of HutC/GntR binding sites in a genome of interest (gray) and for the selection the best HutC/GntR candidate to regulate genes with a specific HutC-like *cis* element pattern (black). For further explanation see text.

of the best candidates. Finally, the searched TF will be validated by experiment to unravel the newly highlighted *cis*/*trans* relationship(s) (steps 13).

Prediction of HutC members target sites and discovery of a new HutC/GntR regulon

By following the above outlined strategy, we have identified all members of the HutC/GntR subfamily in the genomes of *S.coelicolor*. Secondary structure predictions were carried out with all GntR *orfs* present in *S.coelicolor* (57 *orfs*) to define the subfamilies' distribution. Twenty-four HutC/GntR TFs were identified. The search of the HutC TF-binding site consensus sequence (GT-N(1)-TA-N(1)-AC) within the *S.coelicolor* genome revealed 48 sites fitting this sequence, upstream of 62 *orfs*. The deduced or annotated activities of the 62 *orfs* are listed in the table provided in the supplementary data.

Gene annotation of target *orfs* and a literature search identified some genes that are functionally related like *orfs* SCO1390 and 1391 (*crr*, *ptsI*), SCO2905c (*malX2*), SCO2906 (*nagE1*), SCO2907 (*nagE2*) and SCO5841c (*ptsH*). They encode enzyme IIA^{Crr}, enzyme I (EI), enzyme IIB^{Nag}, enzyme IIC^{Nag} and the histidine phosphocarrier protein (HPr) of the phosphotransferase system (PTS) specific for the uptake of *N*-acetylglucosamine (PTS^{Nag}) (20,21; Figure 2A). The internalization of the carbon source by the PTS^{Nag} complex occurs via phosphotransfer from phosphoenolpyruvate (PEP) to enzyme I (EI), to HPr, to IIA^{Crr}, to IIB^{Nag}. IIB^{Nag} in turn phosporylates *N*-acetylglucosamine when the sugar enters the cell through the IIC^{Nag} transport channel (Figure 2B) (20). EI is the only component of the PTS that is missing from the list of predicted target genes however it is encoded immediately downstream of the IIA^{Crr} gene. As gene expression data revealed that *crr* and *ptsI* constitute an operon stimulated by *N*-acetylglucosamine (20), we can presume that

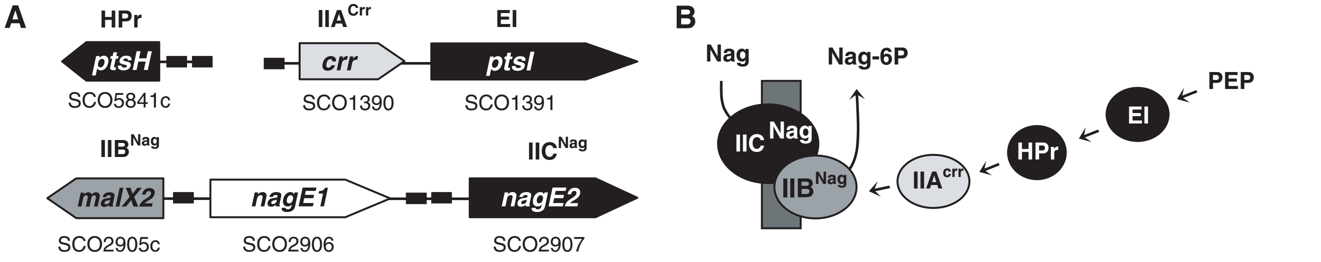

Figure 2. The *N*-acetylglucosamine uptake in *S.coelicolor*. (**A**) Individual *loci* of genes of the PTS^{Nag} regulon. Black boxes indicate positions of the predicted HutC/GntR binding sites. (**B**) Model of *N*-acetylglucosamine uptake via the PTS in *S.coelicolor*. For further explanations see text in the result section.

A **Multiple Alignment**

positions	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Scores
SCO1390	T	G	T	G	G	T	C	T	A	G	A	C	C	T	C	T	14.22
SCO2905c	A	C	T	G	G	T	G	T	A	G	A	C	C	A	G	T	16.46
SCO2907 (1)	A	C	C	G	G	T	C	T	A	C	A	C	C	A	C	T	14.86
SCO2907 (2)	A	G	T	G	G	T	G	T	A	G	A	C	C	A	C	C	15.23
SCO5841c (1)	A	C	T	G	G	T	C	T	A	G	A	C	A	A	C	T	17.10
SCO5841c (2)	A	C	T	G	G	T	C	T	A	G	A	C	A	A	G	A	15.23
Consensus	A	C	T	G	G	T	C	T	A	G	A	C	C	A	C	T	17.74

(I) and (2) are for the first and second predicted TF-binding sites. C, is indicated when the *orf* is on the complementary strand. Scores are in unit of bit.

B **Alignment Matrix**

positions	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	5	0	0	0	0	0	0	0	6	0	6	0	2	5	0	1
C	0	4	1	0	0	0	4	0	0	1	0	6	4	0	4	1
G	0	2	0	6	6	0	2	0	0	5	0	0	0	0	2	0
T	1	0	5	0	0	6	0	6	0	0	0	0	0	1	0	4

Each line displays the number of occurrence of the given nucleotide at the 16 different positions. Numbers enclosed in boxes indicate the number of occurrence of nucleotides that constitute the consensus sequence.

C **“PTScoeli” Weight Matrix**

(Maximum score: 17.74 bits ; Minimun score: -31.20 bits)

positions	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
A	1.10	-1.95	-1.95	-1.95	-1.95	-1.95	-1.95	-1.95	1.27	-1.95	1.27	-1.95	0.25	1.10	-1.95	-0.34
C	-1.95	0.89	-0.34	-1.95	-1.95	-1.95	0.89	-1.95	-1.95	-0.34	-1.95	1.27	0.89	-1.95	0.89	-0.34
G	-1.95	0.25	-1.95	1.27	1.27	-1.95	0.25	-1.95	-1.95	1.10	-1.95	-1.95	-1.95	-1.95	0.25	-1.95
T	-0.34	-1.95	1.10	-1.95	-1.95	1.27	-1.95	1.27	-1.95	-1.95	-1.95	-1.95	-1.95	-0.34	-1.95	0.89

Each line displays the weight of the given nucleotide at the 16 different positions. Scores are in unit of bit. Values enclosed in boxes indicate the scores of nucleotides that constitute the consensus sequence.

Figure 3. Alignment and weight matrix deduced for the PTS^{Nag} cluster.

the HutC-like *cis*-element upstream *crr* also functions for *ptsI*. In addition to this, it should be noted that other genes of a putative *N*-acetylglucosamine regulon also possess HutC-like *cis*-elements. Among these are the four *orfs* that encode chitinases (degradation of the *N*-acetylglucosamine polymer chitin) and the putative NagB enzyme, a glucosamine phosphate isomerase (SCO5236).

Identification of the best HutC-like TF of the PTS^{Nag} regulon in *S.coelicolor*

According to the regulatory codes deduced for the GntR superfamily, the expression of the PTS gene cluster should be controlled by one of the 24 HutC-like members identified in the *S.coelicolor* genome. To build the matrices specific for PTS^{Nag}, we chose the *cis*-elements upstream *crr–ptsI*, between *malX2* and *nagE1*, upstream *nagE2*, and upstream *ptsH* as training set of sequences (Figure 3A). Alignment (Figure 3B) and weight matrices (Figure 3C) were built using the Target Explorer automated tool (22). The matrix with DNA sequences of elements of the PTS has been saved in the public library under the ‘PTScoeli’ denomination. Its maximum and minimum scores are 17.74 and −31.20 bits, respectively. The respective scores of the *cis*-elements used as the training sequences have been reported to the original multiple alignment (Figure 3A). The mean of the training set of sequences is 15.52 bits with an SD of 1.07 bits. To determine the best HutC/GntR candidate, we tested the DNA sequence upstream the 24 HutC/GntR TFs against the ‘PTScoeli’ weight matrix (Figure 3C) under the assumption that regulatory genes often use autoregulation. We obtained a significant score for five out of the 24 HutC/GntR members (Table 1). The best candidate was found to be encoded by *orf* SCO5231 (hereafter designated DasR) with a score of 13.62 bits, obtained for a 16 bp sequence (ACTGGTCTACACCATT) positioned 361 nt upstream of the ATG start codon. Two additional matching sequences with scores of 4.56 and 4.01 bits were identified at 145 and 310 nt from the start codon, respectively. The importance of the gap between the score of *orf* SCO5231 and those obtained for other HutC-like candidates justified that our attention was focused exclusively on DasR to test-specific DNA–protein interactions. This *orf* is a close homolog of the *dasR* gene (*d*eficient in *a*erial mycelium and *s*pore formation) from *Streptomyces griseus* (91% protein identity) (27). In this organism, DasR regulates an adjacent gene cluster encoding a putative sugar ABC transporter (*dasABC*). DasR appears to repress the *dasABC* operon and in an autoregulatory manner its own transcription. The gene organization upstream of *dasR* of *S.coelicolor* is similar to the one observed in *S.griseus* with *orfs* SCO5232, SCO5233 and SCO5234 that are homologous to *dasA*, *dasB* and *dasC*, respectively. However, the levels of conservation within the amino acids primary sequences are drastically eroded (31% for DasAB and 43% for DasC). Expression of the cluster in *S.griseus* has been reported glucose-dependent and results in an ectopic sporulation phenotype (27,28). Hence, it was hypothesized by Seo and co-workers, that the ABC transporter may be involved in the initiation of morphogenesis provoked by the uptake of an extracellular effector, possibly glucose, suggesting an intimate link between sugar uptake/sensing and induction of morphogenesis. According to our prediction DasR should also be considered as the best candidate to control the expression of the PTS^{Nag} regulon.

Table 1. Best HutC/GntR candidates for the PTS^{Nag} regulon

HutC members *(name)*	Position from start codon	HutC promoter matching sequence	Score
SCO5231c (*dasR*)	+361	ACTG**GT**C**TA**C**AC**CATT	13.62
	+145	CTTG**GT**C**TA**GT**C**CATA	4.56
	+310	CTGG**GT**T**T**GA**AC**ACCC	4.01
SCO1262c	+108	ATTG**GT**C**T**TG**AC**CAAG	6.00
SCO4215 (*xlnR*)	−28	AGGT**GT**C**TA**G**A**GAAGG	4.56
SCP2.35	+163	GCCG**GT**C**T**TC**AC**CGCC	4.31
SCO0530c	+135	ATTG**GT**CC**A**TG**C**CACC	4.18

Nucleotides that constitute the HutC/GntR binding site consensus are in bold in the predicted cis sites. DasR, the best candidate, is shaded in gray.

cis/*trans* relationships between DasR and the PTS cluster

DasR of *S.coelicolor* was overproduced as a His-tagged protein in *E.coli* strain Rosetta (DE3) and purified to homogeneity. Electrophoretic mobility shift assays (EMSAs) were performed to investigate whether DasR would bind to the predicted *cis*-acting elements. The DNA probes tested were those selected to build the weight matrix: (i) upstream SCO1390 (*crr* encoding IIA^{Crr}), (ii) upstream SCO2905c (*malX2* encoding IIB^{Nag}), (iii) upstream SCO2907 (*nagE2* encoding EIIC^{Nag}) and (iv) upstream SCO5841c (*ptsH* encoding HPr) (Figure 4A). A single HutC/GntR-binding site was present within probes I and II while two were identified in probes III and IV. Probe V, containing the *cis*-element upstream *crp* of *S.coelicolor* (26), was used as control. A DasR/DNA complex formation was observed for all tested *cis* sites (Figure 4A). Clear signals for complex formation were observed for probes I and II, while signal bands for probes III and IV were more diffuse and indicated the presence of higher molecular weight complexes due to the multiple predicted *cis*-elements. No signal for complex formation was observed with probe V. EMSA performed with the

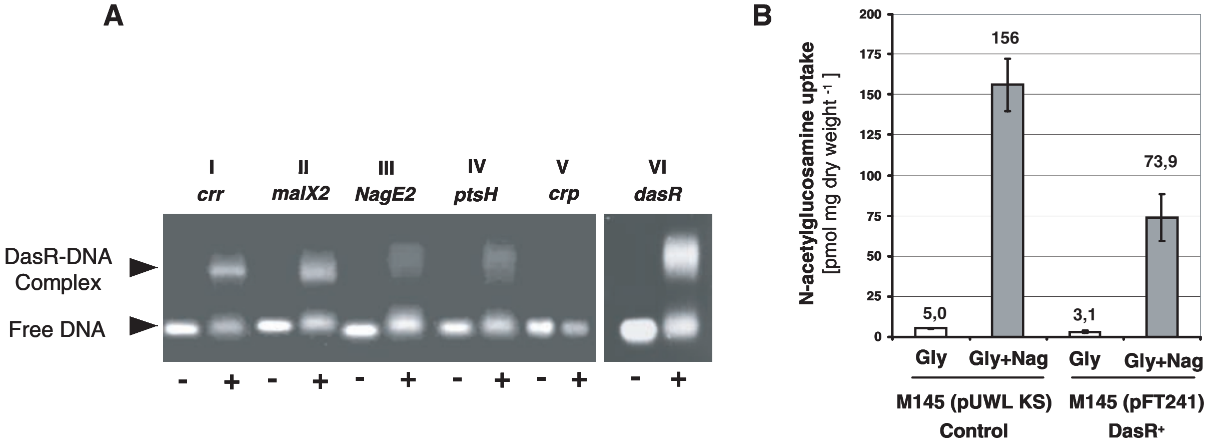

Figure 4. *In vitro* and *in vivo* experimental validation of DasR/PTS^{Nag} *cis*/*trans* relationship. (**A**) Electrophoretic mobility shift assays. A single HutC/GntR-binding site was present within probes I and II while two were identified in probes III and IV. Probe V, containing the predicted *cis*-element of CRP of *S.coelicolor*, was used as control. Probe VI contains the predicted *cis*-element upstream dasR that presented the best score against the ‘PTScoeli’ weight matrix (see Table 1). A DasR/DNA complex formation was observed for all tested HutC-like *cis* sites (probes I to IV). (**B**) *In trans* effect of DasR on *N*-acetylglucosamine transport. Mycelia of M145 (pUWL-SK+), the control, and M145 (pFT241 *dasR*$^{+}$), the DasR overproducing strain, were grown on 50 mM glycerol in the presence and absence of 50 mM *N*-acetylglucosamine. Data points to determine the initial uptake rates were collected as triplicates and the experiment was reproduced three times. Error bars indicate the derived standard deviations.

cis-element upstream of *dasR* (probe VI) also presents a clear signal for complex formation. This result confirms the high score obtained for this *cis*-element against the 'PTScoeli' weight matrix (Table 1), and it is in agreement with those reported for DasR in *S.griseus* (27).

To assess the proposed regulatory role for DasR on the genes of the PTS^{Nag}, plasmid pFT241 ($dasR^+$) was introduced into the wild-type strain M145. Results of the overproduction of DasR in *S.coelicolor* in non-induced (glycerol) or induced (glycerol and *N*-acetylglucosamine) culture conditions are presented in Figure 4B. The $dasR^+$ strain showed a 2-fold reduced induction of *N*-acetylglucosamine transport, an effect that is in agreement with the idea that DasR functions as a negative regulatory element on the PTS regulon. In conclusion, the experiments presented in Figure 4A and B validate our *in silico* approach to discover new *cis/trans* relationships.

DISCUSSION

The work presented here is a direct application of results emerging from a prior bioinformatics analysis of the HTH GntR superfamily of TFs focused on the three components involved in the transcriptional process: (i) the DNA-binding domain, (ii) the effector-binding domain and (iii) the *cis*-acting elements (11). This comparative study was based on the hypothesis that different EBDs impose different sterical constraints on a common type of DBD, which in turn may impose various orientations and presentations of the HTH motif, ultimately reflected by the accommodation of *cis*-acting elements. This hypothesis was demonstrated as for the various subfamilies defined according to the topology of their EBD, we have highlighted different consensus patterns for the recognized *cis*-acting elements. These TF-binding sites consensi were then presented as operator sites prediction tools to search new *cis/trans* relationships in a given genome.

The usefulness of consensus sequences or weight matrices for predicting additional binding sites for well-characterized TFs is not innovative. RegulonDB or DPInteract are actually good tools to refine the prediction of new target genes for a known TF in *E.coli*. Several works report the efficiency of a library of TF-binding sites in conjunction with specific prediction methods such as a 'comparative genomics approach' (29), a 'motif co-occurrence approach' (30) or a 'phylogenetic footprint approach' (31). Nevertheless, the use of these libraries is less appropriate when the aim is to predict completely new *cis/trans* relationships. Using the general *cis* pattern of a large HTH superfamily is unusual assuming that generally the set of training sequences to build the consensus sequence or weight matrix present a large variability that results in an inefficient prediction tool. We have demonstrated that the main advantage in clustering TFs of a large superfamily into different subfamilies, results in a reduction in the variability of the set of DNA training sequences. As a consequence, the global DNA consensus pattern GT-$\text{N}_{(0-15)}$-AC, which was impossible to use as an operator site prediction tool, was replaced by different and more accurate consensi. With the demonstration of specific DNA–protein interactions of DasR and the *cis*-elements upstream genes of the PTS^{Nag} cluster of *S.coelicolor*, we have experimentally validated the proposed regulatory code of the HutC subfamily. The DasR/PTS^{Nag} relationship has been recently confirmed by the knockout of *dasR* in *S.coelicolor*, as the *dasR*-negative strain presents a constitutive expression of genes that constitutes the PTS^{Nag} cluster (our unpublished data). In conclusion, we give a first example showing how starting from an extended comparative of TF superfamilies, one would be able to predict new regulons and limit the search of a specific TF among few candidates out of the hundreds of *orfs* identified as 'regulatory proteins' in a given genome. In the specific case of genes of the PTS^{Nag} cluster, we could measure the efficiency of our prediction approach as the number of the privileged regulatory proteins dropped to the 24 HutC-like TFs from the 673 *orf*s annotated as 'regulatory proteins' in the genome of *S.coelicolor*. With this method, >95% of the regulatory proteins were immediately rejected from the list of the presumed candidates. Most of the time, the selection of TFs candidates would stop at this step and further experimental investigation are required to select the ultimate candidate. In our selected cluster, we could further refine our selection of candidates, as it appeared that the regulator gene itself contained a significant matching sequence detected in its upstream region by the 'PTScoeli' weight matrix.

The prediction approach used in this work was very restrictive, as it was based on a consensus sequence (GT-N(1)-TA-N(1)-AC) and we did not allow mismatches. We imposed such a stringent search accounting for the fact that some positions are more conserved than others, and presumably are more important for the activity of the HutC/GntR site. Using strict consensus impedes the prediction of many target sites as it makes biological sense that TF-binding sites should be variable as regulatory systems can take advantages of this variability to control transcription. In fact, the affinity of DNA–protein interactions required in a specific system does not care about regulatory codes deduced from any compilation. It means that the sequence that should be perfectly recognized by a TF sometimes does not correspond with the biological needs of the organism under a given set of conditions. A perfectly fitted *cis*-acting sequence could, for instance, avoid basal transcription of enzymes required for vital cellular processes. It is important to keep in mind that the sequence searched and predicted may not be experimentally found, not because the prediction method or tools are not valid, but because the expected sequence cannot correspond to what is required *in vivo* under specific environmental conditions. As a result, the list of the HutC-like predicted binding sites could considerably be widened simply by using a weight matrix from the compilation of HutC-like motifs rather than using the consensus sequence.

Interestingly, a recent genome analysis by Studholme *et al.* (32) has presented the identification of several conserved motifs in non-coding regions of *S.coelicolor* genome that could have regulatory functions. One of the proposed *cis* sites were those investigated in this study. Based on previous studies, they suggested that it might be involved in the regulation of *pts* genes and chitinase genes (20,26,33). In addition to their motifs research, they propose one candidate (*chiR*) to control the expression of this carbon uptake/degradation regulon and they invite research groups to confirm their predictions. According to our analysis this regulator is DasR and it remains to be awaited whether ChiR also binds to this target sequence.

Of course, the method is not without pitfalls that are inherent to any search using DNA consensus sequences or weight matrices. This will be the case when the binding site from a member of a well-characterized family does not fit the expected pattern. Such an example is found within the HutC/GntR regulator family, where FarR of *E.coli* binds direct repeat sequences instead of palindromic, inverted repeats (34). Another pitfall could be that the efficiency of a prediction tool will depend on its G+C content and the one of the organism on which the prediction is made. For instance, in our specific case, the HutC/GntR binding site consensus (GT-N(1)-TA-N(1)-AC) was also used to search the *Bacillus subtilis* and *E.coli* genomes for predicted HutC target sites. In *S.coelicolor*, we identified 48 sites. The ratio of the number of predicted *cis* HutC sites versus the number of *trans* HutC members is two (48/24). The value of this ratio largely differs from those obtained with *E.coli* or *B.subtilis* genomes, respectively, 36.3 (218/6) and 46.8 (281/6). These data that could be first interpreted as the average pleiotropicity of HutC members in the studied genomes rather show how the efficiency of a DNA sequence as prediction tool can depend on the G+C content of the organism where the search is made. For instance, the HutC/GntR binding site consensus (G+C ~ 30%) will give more false-positive matching sequences in *E.coli* (G+C ~ 50%) or *B.subtilis* (G+C ~ 40%) than in *Streptomyces* species (G+C ~ 70%) simply because the natural frequency of appearance of the sequence is much more elevated in the first two genera. Hence, each consensus or weight matrix possesses its own efficiency according to the considered microorganism. A further improvement would be the separation of the set of training sequences according to the G+C content of the bacterial strain. This will be feasible when new experimental data will be added to the actual poor source of known TF-binding sites for each HTH subfamily.

Finally, we really want to focus the attention of research groups on the fact that the GntR superfamily is not a unique case in the prokaryotic world where one can find different EBDs fused to a DBD with a common feature. Our bioinformatic analysis will definitely have its validity when similar investigations are made to other TF superfamilies. The accurate TF-binding sites prediction tools that have emerged in conjunction with the experimental validation presented in this work should encourage similar *in silico* approaches in order to reveal new consensi and enrich the list of bacterial *cis*/*trans* regulatory codes. This could be highly facilitated by the creation of databases of *cis*-acting elements with an organized classification according to the various bacterial HTH subfamilies. The ultimate goal is to present, for each sequenced genome, a TF-binding sites cartography and propose the best predicted TF candidate(s) according to the regulatory codes that govern each HTH subfamily.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We thank Marie-Joëlle Virolle, Martin Lee, Colin Smith, Ramon Santamaria, José Manuel Fernández-Abalos, Sonia Rodriguez and Margarita Diaz, Andreas Thomae, Andreas Willimek, Kerstin Mahr, Jean Dusart, Adeline Derouaux, Giorgios Moutzourelis, André Piette, Fabrizio Giannotta, Christine Geron, and Maria Colombo for helpful discussions, and Lisa Rigali for kind support during the preparation of the manuscript. This work was supported in part by the Belgian Program of Interuniversity Poles of Attraction (PAI P5/33), the Walloon Region (convention No. 114930), the Fonds pour la Recherche Fondamentale et Collective, Brussels, Belgium (2.4530.03) and through grants SFB473 and GK805 of the *Deutsche Forschungsgemeinschaft*.

REFERENCES

1. Pabo,C.O. and Sauer,R.T. (1992) Transcription factors: structural families and principles of DNA recognition. *Annu. Rev. Biochem.*, **61**, 1053–1095.
2. Rosinski,J.A. and Atchley,W.R. (1999) Molecular evolution of helix–turn–helix proteins. *J. Mol. Evol.*, **49**, 301–309.
3. Pérez-Rueda,E. and Collado-Vides,J. (2000) The repertoire of DNA-binding transcriptional regulators in *Escherichia coli* K-12. *Nucleic Acids Res.*, **28**, 1838–1847.
4. Hulo,N., Sigrist,C.J.; Le Saux,V., Langendijk-Genevaux,P.S., Bordoli,L., Gattiker,A., De Castro,E., Bucher,P. and Bairoch,A. (2004) Recent improvements to the PROSITE database. *Nucleic Acids Res.*, **32**, D134–D137.
5. Brown,N.L., Stoyanov,J.V., Kidd,S.P. and Hobman,J.L. (2003) The MerR family of transcriptional regulators. *FEMS Microbiol. Rev.*, **27**,145–163.
6. Busenlehner,L.S., Pennella,M.A. and Giedroc,D.P. (2003) The SmtB/ArsR family of metalloregulatory transcriptional repressors: structural insights into prokaryotic metal resistance. *FEMS Microbiol. Rev.*, **27**, 131–143.
7. Korner,H., Sofia,H.J. and Zumft,W.G. (2003) Phylogeny of the bacterial superfamily of Crp-Fnr transcription regulators: exploiting the metabolic spectrum by controlling alternative gene programs. *FEMS Microbiol. Rev.*, **27**, 559–92.
8. Weickert,M.J. and Adhya,S. (1992) A family of bacterial regulators homologous to Gal and Lac repressors. *J. Biol. Chem.*, **267**, 15869–15874.
9. Haydon,D.J. and Guest,J.R. (1991) A new family of bacterial regulatory proteins. *FEMS Microbiol Lett.*, **63**, 291–295.
10. Rigali,S., Derouaux,A., Giannotta,F. and Dusart,J. (2002) Subdivision of the helix–turn–helix GntR family of bacterial regulators in the FadR, HutC, MocR, and YtrA subfamilies. *J. Biol. Chem.*, **277**, 12507–12515.
11. Schell,M.A. (1993) Molecular biology of the LysR family of transcriptional regulators. *Annu. Rev. Microbiol.*, **47**, 597–626.
12. Mota,L.J., Tavares,P. and Sa-Nogueira,I. (1999) Mode of action of AraR, the key regulator of L-arabinose metabolism in *Bacillus subtilis*. *Mol. Microbiol.*, **33**, 476–489.
13. Lee,M.H., Scherer,M., Rigali,S. and Golden,J.W. (2003) PlmA, a new member of the GntR family, has plasmid maintenance functions in *Anabaena* sp. strain PCC 7120. *J. Bacteriol.*, **185**, 4315–4325.
14. Robison,K., McGuire,A.M. and Church,G.M. (1998) A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. *J. Mol. Biol.*, **284**, 241–254.
15. Salgado,H., Gama-Castro,S., Martinez-Antonio,A., Diaz-Peredo,E., Sanchez-Solano,F., Peralta-Gil,M., Garcia-Alonso,D., Jimenez-Jacinto,V., Santos-Zavaleta,A., Bonavides-Martinez,C. and Collado-Vides,J. (2004) RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res.*, **32**, D303–D306.
16. Stormo,G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.
17. Stormo,G.D. and Tan,K. (2002) Mining genome databases to identify and understand new gene regulatory systems. *Curr. Opin. Microbiol.*, **5**, 149–153.
18. Aravind,L. and Anantharaman,V. (2003) HutC/FarR-like bacterial transcription factors of the GntR family contain a small molecule-binding domain of the chorismate lyase fold. *FEMS Microbiol. Lett.*, **222**, 17–23.

19. Bentley,S.D., Chater,K.F., Cerdeno-Tarraga,A.M., Challis,G.L., Thomson,N.R., James,K.D., Harris,D.E., Quail,M.A., Kieser,H., Harper,D. *et al*. (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature*, **417**, 141–147.
20. Nothaft,H., Dresel,D., Willimek,A., Mahr,K., Niederweis,M. and Titgemeyer,F. (2003) The phosphotransferase system of *Streptomyces coelicolor* is biased for N-acetylglucosamine metabolism. *J. Bacteriol*., **185**, 7019–7023.
21. Parche,S., Nothaft,H., Kamionka,A. and Titgemeyer,F. (2000) Sugar uptake and utilisation in *Streptomyces coelicolor*: a PTS view to the genome. *Antonie Van Leeuwenhoek*, **783**, 243–251.
22. Sosinsky,A., Bonin,C.P., Mann,R.S. and Honig,B. (2003) Target Explorer: an automated tool for the identification of new target genes for a specified set of transcription factors. *Nucleic Acids Res*., **31**, 3589–3592.
23. Hertz,G.Z.and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.
24. Parche,S., Schmid,R. and Titgemeyer,F. (1999) The phosphotransferase system (PTS) of *Streptomyces coelicolor* identification and biochemical analysis of a histidine phosphocarrier protein HPr encoded by the gene *ptsH*. *Eur. J. Biochem*., **265**, 308–317.
25. Studier,F.W., Rosenberg,A.H., Dunn,J.J. and Dubendorff,J.W. (1990) Use of T7 RNA polymerase to direct expression of cloned genes. *Methods Enzymol*., **185**, 60–89.
26. Derouaux,A., Halici,S., Nothaft,H., Neutelings,T., Moutzourelis,G., Dusart,J., Titgemeyer F. and Rigali,S. (2004) Deletion of a cyclic AMP receptor protein homologue diminishes germination and affects morphological development of *Streptomyces coelicolor*. *J Bacteriol*., **186**, 1893–1897.
27. Seo,J.W., Ohnishi,Y., Hirata,A. and Horinouchi,S. (2002) ATP-binding cassette transport system involved in regulation of morphological differentiation in response to glucose in *Streptomyces griseus*. *J. Bacteriol*., **184**, 91–103.
28. Ohnishi,Y., Seo,J.W. and Horinouchi,S. (2002) Deprogrammed sporulation in *Streptomyces*. *FEMS Microbiol. Lett*., **29**, 1–7.
29. Tan,K., Moreno-Hagelsieb,G., Collado-Vides,J. and Stormo,G.D. (2001) A comparative genomics approach to prediction of new members of regulons. *Genome Res*. **11**, 566–584.
30. Bulyk,M.L., McGuire,A.M., Masuda,N. and Church,G.M. (2004) A motif co-occurrence approach for genome-wide prediction of transcription-factor-binding sites in *Escherichia coli*. *Genome Res*., **14**, 201–208.
31. McCue,L., Thompson,W., Carmack,C., Ryan,M.P., Liu,J.S., Derbyshire,V. and Lawrence,C.E. (2001) Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes. *Nucleic Acids Res*., **29**, 774–782.
32. Studholme D.J., Bentley S.D. and Kormanec J. (2004) Bioinformatic identification of novel regulatory DNA sequence motifs in *Streptomyces coelicolor*. *BMC Microbiol*., **4**, 14.
33. Ni,X. and Westpheling,J. (1997) Direct repeat sequences in the *Streptomyces* chitinase-63 promoter direct both glucose repression and chitin induction. *Proc. Natl Acad. Sci. USA*, **94**, 13116–13121.
34. Quail,M.A., Dempsey,C.E. and Guest,J.R. (1994) Identification of a fatty acyl responsive regulator (FarR) in *Escherichia coli*. *FEBS Lett*., **356**, 183–187.