BMC Microbiology

Research article

Bioinformatic identification of novel regulatory DNA sequence motifs in *Streptomyces coelicolor*

David J Studholme*[1], Stephen D Bentley[1] and Jan Kormanec[2]

Address: [1]Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK and [2]Institute of Molecular Biology, Center of Excellence for Molecular Medicine, Slovak Academy of Sciences, Dubravska cesta 21, 845 51 Bratislava, Slovak Republic

Email: David J Studholme* - ds2@sanger.ac.uk; Stephen D Bentley - sdb@sanger.ac.uk; Jan Kormanec - umbijkor@savba.savba.sk

* Corresponding author

Published: 08 April 2004

BMC Microbiology 2004, **4**:14

Received: 16 October 2003
Accepted: 08 April 2004

This article is available from: http://www.biomedcentral.com/1471-2180/4/14

Abstract

Background: *Streptomyces coelicolor* is a bacterium with a vast repertoire of metabolic functions and complex systems of cellular development. Its genome sequence is rich in genes that encode regulatory proteins to control these processes in response to its changing environment. We wished to apply a recently published bioinformatic method for identifying novel regulatory sequence signals to gain new insights into regulation in *S. coelicolor*.

Results: The method involved production of position-specific weight matrices from alignments of over-represented words of DNA sequence. We generated 2497 weight matrices, each representing a candidate regulatory DNA sequence motif. We scanned the genome sequence of *S. coelicolor* against each of these matrices. A DNA sequence motif represented by one of the matrices was found preferentially in non-coding sequences immediately upstream of genes involved in polysaccharide degradation, including several that encode chitinases. This motif (TGGTCTAGACCA) was also found upstream of genes encoding components of the phosphoenolpyruvate phosphotransfer system (PTS). We hypothesise that this DNA sequence motif represents a regulatory element that is responsive to availability of carbon-sources.

Other motifs of potential biological significance were found upstream of genes implicated in secondary metabolism (TTAGGTtAGgCTaACCTAA), sigma factors ($TGACN_{19}TGAC$), DNA replication and repair ($ttgtCAGTGN_{13}TGGA$), nucleotide conversions (CTACgcNCGTAG), and ArsR ($TCAGN_{12}TCAG$). A motif found upstream of genes involved in chromosome replication ($TGTCagtgcN_7Tagg$) was similar to a previously described motif found in UV-responsive promoters.

Conclusions: We successfully applied a recently published *in silico* method to identify conserved sequence motifs in *S. coelicolor* that may be biologically significant as regulatory elements. Our data are broadly consistent with and further extend data from previously published studies. We invite experimental testing of our hypotheses *in vitro* and *in vivo*.

Background

The complete functional information encoded in a genome consists not only of the genes, encoding proteins and RNAs, but also structural and regulatory elements. Complete genome sequences are generated much faster than regulatory sites can be determined by experiment,

and so there is a need for computational prediction and detection of regulatory elements in complete genomes.

In some cases, a set of co-regulated genes (*i.e.* a regulon) is known, for example as the result of micro-array experiments. Several algorithms are available to discover sequence motifs that occur more frequently in a set of co-regulated genes compared against the background sequence [*e.g.* [1-6]]. In the absence of experimental evidence for co-regulation, an alternative approach is to compare the upstream regions of orthologous genes from different species [7-9].

Successful as these procedures have been, they also have limitations. For example, mRNA micro-array experiments will only find those regulons that are responsive to the laboratory conditions tested, and interspecies comparison is limited by the availability of species separated by an appropriate phylogenetic distance.

Several recent studies have addressed the problem of discovering binding sites for regulatory proteins in the absence of experimental data about co-regulation [*e.g.* [10,11]]. Again these methods are based on identifying short DNA sequences that occur more frequently in a subset of sequences than would be expected from a random sample of the background distribution. Li *et al*. [11] carried out a comprehensive analysis of the complete *Escherichia coli* K12 genome sequence and discovered many new potential regulatory elements as well as recovering previously known binding sites for 37 regulatory proteins. Subsequently, Mwangi and Siggia [12] applied the same technique, with some refinements, to the complete genome of the Gram-positive bacterium *Bacillus subtilis*.

Streptomyces species are Gram-positive soil bacteria that undergo a complex cycle of morphological differentiation, which culminates in the spore-bearing aerial hyphae on mycelial colonies. This process of morphological differentiation is also accompanied by physiological differentiation, characterised by production of biologically active secondary metabolites. In fact *Streptomyces* species produce the majority of known antibiotics [13]. The huge metabolic repertoire and complex developmental cycle is the product of one of the largest known prokaryotic genomes. The genetically best-known representative of the genus, *S. coelicolor* A3(2), possesses an 8.67 Mb linear chromosome [14]. Its complex genome is particularly rich in regulatory proteins (12.3% from 7825 predicted genes) including about 65 sigma factors [14]. We applied a similar method to that of Li *et al*. [11] and Mwangi and Siggia [12] to try to discover novel regulatory motifs in *Streptomyces coelicolor*.

Results and discussion

The non-coding intergenic regions of the *S. coelicolor* genome were scanned for significantly over-represented dyads of the form W_1NxW_2 (see Methods for more details) where W_1 and W_2 each represent a sequence of 4 nucleotides and Nx represents a sequence of between 4 and 20 nucleotides long. This resulted in identification of 2770 dyads that were statistically over-represented. These are listed in additional file 10 [significantly_overrepresented_dyads.txt]. The 2770 dyads were grouped into 2497 clusters, each of which was used to generate a position-specific weighted scoring matrix (PSWM). Essentially, each PSWM represented a statistical model of a DNA sequence motif, which summarises the relative frequencies of each of the four nucleotides at each position. Each of the matrices is listed in additional file 11 [matrices.txt].

All the intergenic sequences in the *S. coelicolor* genome were scanned against each of the matrices to find high-scoring matches to the DNA sequence motifs represented by each matrix. Thus we identified 2497 sets of genes that share common upstream DNA sequence motifs. Given this large quantity of data, the challenge was to determine which of these sets of genes sharing common upstream sequence motifs represented biologically significant regulons.

One way to address this issue was to search the data for sets of genes that are functionally coherent; that is to find those sets of genes in which the number of members associated with a particular biological function is significantly larger than that which could be explained by chance. As an indicator of biological function, we made use of a protein classification scheme [15] (see additional file 12 [classification_scheme.txt]) based on that originally created for *E. coli* in the EcoCyc database [16,17]. For every set of genes sharing an upstream DNA sequence motif, each gene was assigned to one of the 181 functional classes. To determine whether any functional class was statistically significantly over-represented in the set of genes, a P value was calculated. The P value represented the probability of obtaining the observed number of genes (or more) belonging to that functional class within a set. This P value indicates the statistical significance of functional coherence within a set of genes.

We chose a threshold P value that was equal to one divided by the product of the number of matrices and the number of categories considered: $1 / (2497 * 181) = 2.21 \times 10^{-6}$. Thus by chance alone for the complete analysis, we would expect to find less than one set of genes in which the P value is less than this threshold. In fact, we found 11 sets of genes sharing common upstream sequence motifs that contained over-represented functional categories

Table 1:

Matrix	Protein class		Number of ORFs in this set	Number of ORFs in this set belonging to this protein class	P value	Consensus sequence
2083	2.1.3	Degradation of polysaccharides	54	10	7.557e-10	See Figure 2A
2318	2.2.3	DNA – replication, repair, restriction / modification	21	6	7.76e-08	See Figure 2B
1744	1.2.1	Chromosome replication	24	3	2.12e-06	See Figure 2C
1909	4.1.7	Gram +ve exported / lipoprotein	106	22	9.24e-08	See Figure 2D
46	6.2.1	sigma factor	116	10	6.60e-08	See Figure 2E
2034	6.3.13	ArsR	45	4	1.89e-06	See Figure 2F
1853	3.3.11	Nucleotide interconversions	46	5	4.09e-07	See Figure 2G
363	3.8.0	Secondary metabolism	9	5	4.89e-07	See Figure 2H
571	3.8.0	Secondary metabolism	10	5	9.621e-07	See Figure 2H
293	3.8.0	Secondary metabolism	10	5	9.62e-07	See Figure 2H
153	3.8.0	Secondary metabolism	18	6	1.31e-06	See Figure 2H

Table 1. Position-specific weight matrices (PSWMs) that represent DNA sequence motifs shared by functionally coherent sets of genes in *Streptomyces coelicolor*. A library of 2497 matrices was generated from alignments of over-represented DNA sequence dyads as described in the Methods section. Each matrix is essentially a statistical model of a DNA sequence motif [58]. The non-coding regions of the *S. coelicolor* genome were searced against the matrices to find matches to each of the sequence motifs. The scanning method assigneda score (maximum 100) to each match site. The minimum score threshold was chosen as 80. For each matrix, we recorded the number of genes whose upstream region contains at least one match site. We also recorded the number of those genes belonging to each functional category in the protein classification scheme, and calculated a P value to determine whether that functional category was significantly over-represented.

(Table 1 and Figure 2) plus several additional motifs associated with category 5.1.4 "Transposon/insertion element-related functions". The motifs associated with category 5.1.4 were excluded from further consideration as the aim of this study was to identify potential regulatory elements, not repeats associated with mobile elements.

We performed a negative control experiment whereby we randomly shuffled the orders of the columns of each of the 2497 matrices and repeated the scans. After excluding several motifs associated with category 5.1.4 "Transposon/insertion element-related functions", none of the matrices yielded sets of target genes that satisfied the criteria for functional coherence.

It is important to note that for a set of genes to be functionally coherent, it is not necessary that the majority of members of the set belong to a particular category; it is only necessary that the number of members belonging to a particular functional class is greater than could be reasonably expected by chance.

A motif associated with degradation of polysaccharides and sugar transport

One of the matrices (matrix 2083), when used to search against the *S. coelicolor* genome, yielded matches upstream of sets of genes in which the functional class 2.1.3 "Degradation of polysaccharides" were over-represented (Table 2, additional file 1 [table2.pdf]). Many of the sites were also conserved in the close relative *S. avermitilis* (Table 10, additional file 9 [table10.pdf]). Scanning the genome sequence of *S. coelicolor* against matrix 2083 yielded intergenic match sites upstream of 54 genes, including nine that encode chitinases (SCO1429, SCO2503, SCO5003, SCO5376, SCO5673, SCO5954, SCO6012, SCO7225 and SCO7263) and two that encode chitin-binding proteins (SCO0481 and SCO2833). The intergenic regions upstream of SCO2503, SCO6012, SCO7225 and SCO7263 each contain two sites (*i.e.* tandem repeats).

Chitinases (EC 3.2.1.14) are a group of glycosylhydrolases that hydrolyse chitin, a polymer of N-acetylglucosamine linked by beta-1, 4-bonds. Streptomycetes are the main decomposers of chitin, the second most abundant polysaccharide in soil. The complete genome of *S. coelicolor* [14] encodes 11 chitinases [18].

Several studies have demonstrated that *Streptomyces* chitinase genes are induced in the presence of chitin and repressed in the presence of glucose plus chitin [19-22]. A direct repeat of the sequence TGGTCCAGACCT, similar to that of our motif, was shown [23] to be involved in the both chitin-induction and in glucose-repression of *chi63* encoding chitinase C (SCO5376). A similar motif has been identified upstream of several other chitinases in *S. coelicolor* [23-25].

From the results of our searches against matrix 2083, it was apparent that several genes had direct repeats of the motif in their upstream non-coding regions. These included genes encoding chitinases (SCO2503, SCO6012, SCO7225 and SCO7263), phosphoenolpyruvate phosphotransfer transport system (PTS) components

(SCO2907 and SCO5841), a hydrolase (SCO6032), amongst others (Table 2, additional file 1 [table2.pdf]). The distance between repeated instances of the motif appeared to be quite variable.

The identity of the transcription factor(s) mediating regulation of *chi63* through this DNA element is not known. In *S. lividans*, glucose repression of chitinase production involves glucose kinase (encoded by *glkA*) [26,27]. However, it has been demonstrated that *glkA* is not required for glucose repression of *chi63* in *S. coelicolor* [28].

In *S. lividans*, regulation of chitinases also requires Reg1, the 345-amino acid product of the gene *reg1* [29,30]. Reg1 (SwissProt P72469) is a member of the LacI/GalR family of transcriptional regulators and contains a helix-turn-helix motif at its N terminus. Reg1 is 95% identical to MalR, the repressor of *malE* in *S. coelicolor*, and *reg1* shares synteny with *malR*; so Reg1 can be considered to be the orthologue of MalR [30]. Two Reg1 (MalR) binding sites upstream of *malE* have been identified in *S. lividans* by DNA footprinting [31]. These operator sequences share no obvious similarity with the motif represented by matrix 2083. Therefore it is unlikely that this motif represents the Reg1/MalR binding signal.

Interestingly, the intergenic sites matched by matrix 2083 also included several that are upstream of genes encoding components of the PTS. The main function of the PTS is to carry out the fundamental process of substrate-level phosphorylation, whereby the transport and activation of sugar substrates are thermodynamically coupled to dephosphorylation of the glycolytic intermediate phosphoenolpyruvate. The phosphoryl relay proceeds sequentially from phosphoenolpyruvate to enzyme I (EI), HPr, enzyme II (EII), and finally to the incoming sugar that is transported across the membrane and concomitantly phosphorylated by EII. Additionally, in Gram-negative and low G+C Gram-positive bacteria, PTS components have been shown to participate in signal transduction, chemotaxis, and the regulation of some key physiological processes, *e.g.*, carbohydrate transport, catabolite repression, carbon storage, and coordination of carbon and nitrogen metabolism [32,33]. Complete genome sequencing confirmed the presence of the general PTS components enzyme I (*ptsI*), HPr (*ptsH*) and enzyme IIA-Crr (*crr*) in *S. coelicolor* [14,34-37]. Also there are several genes that may encode four sugar-specific PTS permeases. Several studies have examined the PTS in *Streptomyces* species and demonstrated its activity [*e.g.* [34-38]].

Sites matching matrix 2083 were found immediately upstream of several *S. coelicolor* genes that encode PTS EII components (SCO1390, SCO2906 and SCO2907) and *ptsH* (SCO5841) that encodes HPr. The co-occurrence of similar DNA elements upstream of genes encoding glucose-repressible chitinases and upstream of PTS genes, suggests that a regulatory mechanism may be common to both sets of genes. More specifically, an unidentified transcription factor that is presumed to mediate chitin-induction and glucose repression of *chi63* may also play a role in regulation of the PTS in *S. coelicolor*.

In addition to those associated with chitinases and PTS components, a number of other sites matching matrix 2083 (Table 2, additional file 1 [table2.pdf]) were also found upstream of genes involved in carbohydrate metabolism, including acetyl-coenzyme A synthetase (SCO3563), a putative glucosamine phosphate isomerase (SCO5236) and several putative sugar binding proteins (SCO0531, SCO2946 and SCO5232). It seems plausible that these might also share a common regulatory system with the PTS and chitinases.

Conserved DNA sequence motifs associated with DNA-replication and repair

Of the 21 ORFs having an upstream match to matrix 2318, six belong to functional category 2.2.3 "DNA – replication, repair, restriction/modification" (Table 3, additional file 2 [table3.pdf]). These included genes predicted to encode an ABC excision nuclease (SCO1966), DNA helicase (SCO5761), DNA polymerase (SCO3541 and SCO6084), a DNA gyrase (SCO5836) and a putative primosomal protein (SCO1475).

A sequence motif represented by matrix 1744 was found in intergenic regions upstream of three genes classified as belonging to category 1.2.1 "Chromosome replication" (Table 4, additional file 3 [table4.pdf]). The genes encode a putative replicative DNA helicase (SCO3911), a DNA polymerase subunit (SCO2064) and recombinase RecA (SCO5769). Matches also occurred upstream of other genes implicated in nucleic acid metabolism including ABC excision nuclease subunits (SCO1953, SCO1958 and SCO1966), a helicase (SCO6262) and uracil-DNA glycosylase (SCO1343). On further examination of the matrix 1744 match site upstream of *S. coelicolor recA*, it became apparent that this site overlaps the previously described *recA* promoter [39]. The *recA* promoters of *Streptomyces* and *Mycobacterium* species represent a distinct class of promoter whose consensus sequence is TTGTCAGTGGCN$_6$TAGggT, and which are probably responsive to ultra-violet (UV) light [39]. Clearly the motif represented by matrix 1744 (TGTCagtgcN$_7$TAgg) is related to this promoter sequence consensus.

A conserved DNA sequence motif associated with sigma factors

Bacterial sigma (σ) factors confer upon RNA polymerase the ability to recognise promoter sequences. *Streptomyces*

coelicolor encodes about 65 sigma factors, each perhaps having different promoter-specificities. One important family of sigma factors are the extracytoplasmic function (ECF) sigma factors, of which *S. coelicolor* encodes about 50 [40]. We found matches to matrix 46 in intergenic regions upstream of 116 genes. Of the 66 *S. coelicolor* genes belonging to class 6.2.1 "Sigma factor", ten had matches to matrix 46 in their upstream intergenic regions (SCO0038, SCO0803, SCO2742, SCO3356, SCO4960, SCO7104, SCO7105, SCO7112, SCO7192 and SCO5216). A match was also found upstream of SCO5216, annotated as a sigma factor (Table 5, additional file 4 [table5.pdf]). All eleven of these gene products are predicted to belong to the ECF family of sigma factors, though SCO0038 is only fragment, not a complete sequence. It is possible that this motif, $TGACN_{19}TGAC$, may be involved in some hitherto unknown mechanism of regulation shared by several ECF sigma factors in *S. coelicolor*.

A conserved DNA sequence motif associated with exported proteins

Matrix 1909 matched 106 intergenic sites, 22 of which were upstream of genes belonging to functional class 4.1.7 "Gram positive exported / lipoprotein" (Table 6, additional file 5 [table6.pdf]). The putative products of these 22 genes included many proteins of unknown function predicted to be secreted across the cell membrane (e.g. SCO2741, SCO2001 and SCO1650). Also this $GAACN_{19}GTTG$ motif was found upstream of genes for six secreted peptidases. According to the MEROPS classification [41], these belong to peptidase family S8 (SCO0432, SCO1741), family M1 (SCO7605), family M6 (SCO2920), family M16 (SCO5837) and family M23 (SCO3368). A homologue of SCO0432 (38% sequence identity) cloned and purified from *S. albogriseolus* has been shown to be secreted and to have subtilisin-like endopeptidase activity [42]. Sigma factors belonging to the ECF family characteristically recognise an 'AAC' motif in the -35 region [40]. We hypothesise that this $GAACN_{19}GTTG$ motif may represent a class of promoters targeted by an as yet unidentified ECF sigma factor that regulates expression of some or all of these extracellular proteins.

A conserved DNA sequence motif found upstream of arsR *homologues*

Members of the ArsR family are metalloregulatory transcriptional repressors, which repress expression of operons involved in response to high concentrations of heavy metal ions [43]. Binding of metal ions to the ArsR protein leads to derepression of the target operons. For example, an ArsR-homologue, SrnR, was shown to mediate nickel-responsive transcriptional repression of superoxide dismutase in *S. griseus* [44].

Fifteen genes in *S. coelicolor* are predicted to encode ArsR-like transcriptional regulators [15]. Of these fifteen, four (SCO0875, SCO3699, SCO6808, and SCO6836) have an upstream match to the motif $TCAGN_{12}TCAG$, represented by matrix 2034 (Table 7, additional file 6 [table6.pdf]). In fact SCO6836 has two upstream matches and SCO3699 has three upstream matches. It is possible that this motif represents a regulatory element. None of the other genes with upstream matches are obviously implicated in metal ion stress, so it is unlikely that this is a motif associated generally with metal ion stress response. Also, these matches are not conserved upstream of ArsR homologues in *S. avermitilis*.

A conserved DNA sequence motif found upstream of genes involved in nucleotide metabolism

The genome of *S. coelicolor* encodes 25 proteins belonging to category 3.3.11 "Nucleotide interconversions" (Table 8, additional file 7 [table8.pdf]). In the upstream intergenic regions of five of these, there are matches to the near-palindromic motif CTACgcNCGTAG represented by matrix 1853 (SCO1776, SCO2015, SCO4901, SCO4914 and SCO4917). Furthermore, a match is found upstream of SCO4886. SCO4886 may be the first gene in an operon also containing SCO4889, which encodes a putative cytidine deaminase. A match is also found upstream of SCO4152, encoding a secreted nucleotidase and *thiC* (SCO3938), which is involved in biosynthesis of the nucleotide-derived cofactor thiamine. In many other bacteria, including the Gram-positive *Bacillus subtilis* and *Mycobacterium tuberculosis*, immediately upstream of *thiC* is a highly conserved motif called the *thi* box [45], which is characteristic of the so-called TPP riboswitch [46]. However, according to the Rfam database [47], there is no recognisable TPP riboswitch in the vicinity of *thiC* in *S. coelicolor*.

This motif seems to be a good candidate regulatory element and it would be intriguing to test whether it is involved in regulation of nucleotide metabolism and possibly associated secondary metabolic pathways.

A conserved DNA sequence motif found upstream of genes involved in secondary metabolism

Actinomycetes are famous as producers of antibiotics and other useful products for biotechnology [48]. In the *S. coelicolor* genome, the products of 165 genes were classified as belonging to category 3.8.0 "Secondary metabolism" (Table 9, additional file 8 [table9.pdf]). Three matches to the near-palindrome TTAGGTtAGgCTaACCTAA occur in intergenic regions within a cluster of these genes. This motif was represented by several related matrices (363, 571, 293 and 153). One match falls between the divergently transcribed SCO0489 and SCO0490, the second falls between the divergently transcribed SCO0498 and

SCO0499, and the third falls upstream of SCO0495. It seems plausible that this motif may be involved in coordinated regulation some or all of the genes in this cluster. Their precise function of is unknown, but since these enzyme-encoding genes are clustered together on the chromosome it is possible that they may encode components of a single metabolic pathway. An additional incidence of the motif occurs upstream of a further gene implicated in secondary metabolism (SCO2782).

Distribution of motifs in coding and non-coding DNA

A biologically significant regulatory sequence motif would be expected to occur more frequently in non-coding rather than coding regions of a bacterial genome sequence. Therefore, we investigated whether there was such a bias in the distribution of matches to each matrix in the complete genome of *S. coelicolor* (Figure 1A). We scanned the complete genome sequence of *S. coelicolor* against the matrix using a range of different threshold scores. For each threshold used, we counted the number of matches occurring in intergenic regions and those occurring within coding regions, and calculated the ratio between the two counts. About 5% of the *S. coelicolor* genome is predicted to be non-coding [14]. When low threshold scores were used, for most matrices about 10% of the matches fell within intergenic regions. However, when the searches were made more stringent by raising the threshold, the ratio clearly changed and there was a clear bias towards non-coding regions such that using a threshold of 90, over half the matches occur in the 5% of the *S. coelicolor* genome that is non-coding. The distribution of matches to matrix 363 was exceptionally biased towards non-coding regions, with about 90% of the matches occurring in non-coding DNA even when the threshold was lowered to 70.

A similar pattern was observed when the analysis was repeated using the closely related *S. avermitilis* genome [49] (Figure 1B) for most of the matrices. However, the bias was much smaller for matrix 46, and apparently absent for matrix 1909. This suggested that these motifs are probably not biologically significant in *S. avermitilis*. If matrix 1909 does indeed represent a recognition site for an ECF sigma factor as suggested above, then presumably the sigma factor is not conserved in *S. avermitilis*.

When the same matrices were used to scan the genome of the more distantly related *Mycobacterium tuberculosis* [50] (Figure 1C), the bias towards non-coding DNA was much less pronounced. However, matches to matrix 363 and matrix 1744 showed a significant bias towards non-coding DNA in *M. tuberculosis* as well as in *Streptomyces* species, suggesting that they might be biologically significant in a broader range of organisms. Consistent with this observation, it has previously been reported that the UV-responsive promoter motif represented by matrix 1744 is conserved in mycobacteria as well as streptomycetes [39]. In the genome of the very distantly related *E. coli* [51], there was little or no detectable bias towards occurrence in non-coding DNA (Figure 1D).

Conclusions

Historically, the study of regulation in *Streptomyces* species has focussed on antibiotic production and cellular differentiation and relatively few regulons have been confirmed. We successfully applied a bioinformatic technique [11,12], which had been developed using the model bacteria *E. coli* and *B. subtilis*, to identify conserved DNA motifs in the actinomycete *S. coelicolor*. This resulted in a large number of sets of genes sharing common upstream DNA sequence motifs. These sets could be considered as 'putative regulons', but it is likely that many or most of the motifs would not in fact be biologically significant. Therefore, we selected a small subset of the 'putative regulons' that were functionally coherent (Table 1 and and Figure 2), and therefore might genuinely represent sets of genes that share regulatory DNA elements in common.

One of the functionally coherent putative regulons included genes encoding PTS components, chitinases and other carbohydrate degradation and transport functions. Members of this putative regulon shared an upstream sequence motif (TGGTCTAGACCA). Since the roles of glucose kinase and Reg1 on chitinase regulation are indirect, the mechanism of regulation of *S. coelicolor* chitinases remains unknown. The presence of a sequence motif that is shared by several chitinases and other genes of carbohydrate metabolism is intriguing. One possibility is that the shared sequence motif represents the target site for a DNA-binding transcription factor. A candidate regulatory factor is ChiR [52], which has been shown to be involved in activation of *chiC* expression in *S. coelicolor*. ChiR belongs to the two-component response regulator family [53] and contains a good candidate DNA-binding domain. However, the target DNA sequence for ChiR is unknown, as no direct DNA-binding has yet been demonstrated [52].

We discovered a motif found upstream of genes involved in replication (Table 4, additional file 3 table4.pdf]), which was similar to that of a previously described class of UV-inducible promoters [39]. Also we found conserved sequence motifs associated with ECF sigma factors, ArsR-like transcriptional regulators, nucleotide metabolism, and secondary metabolism. We hope that these *in silico* predictions will provide a useful starting point for future experimental work (such as micro-array studies) on important regulatory systems in *Streptomyces*.

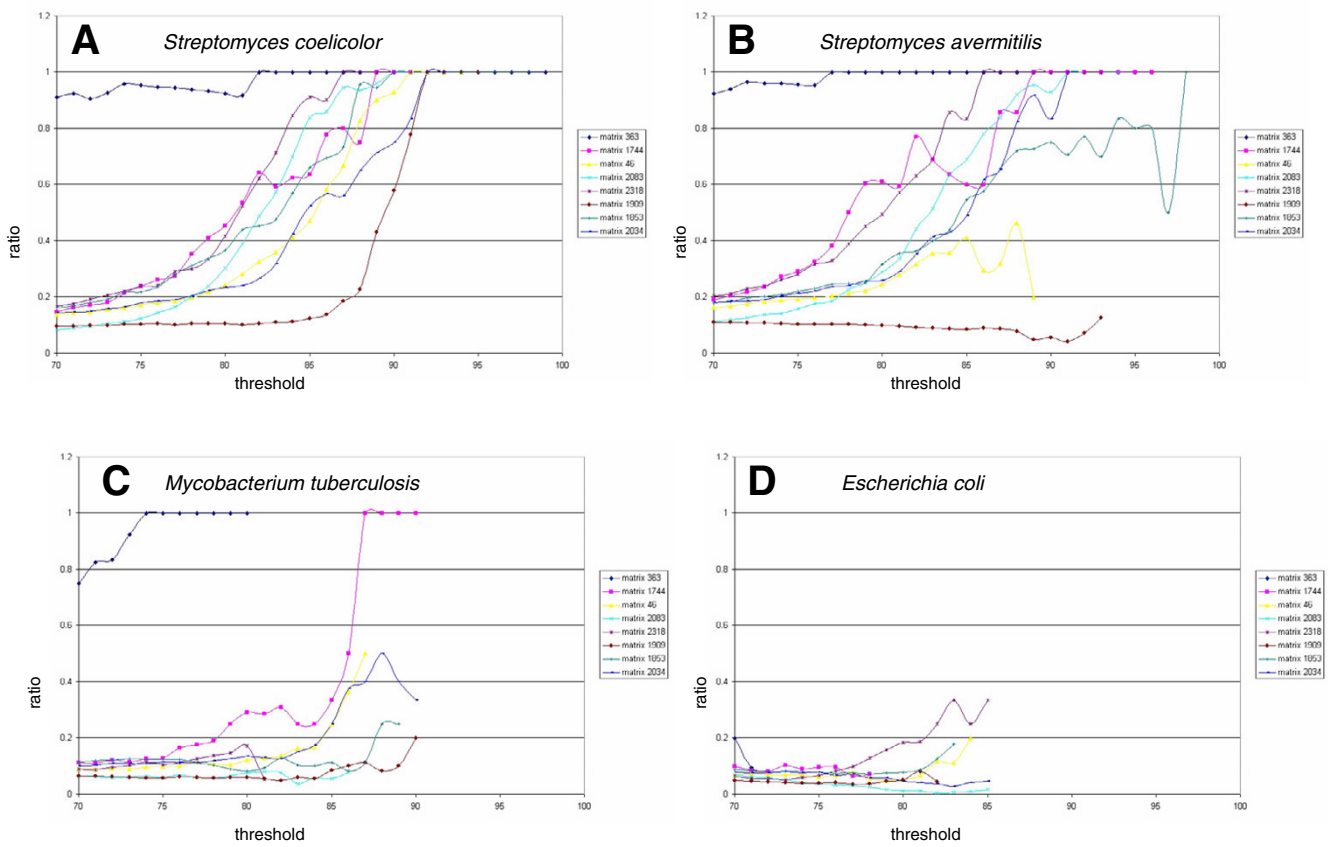

Figure 1
Occurrence of sequence motifs in coding and non-coding DNA. The genomes of *Streptomyces coelicolor* (A), *S. avermitilis* (B), *Mycobacterium tuberculosis* (C) and *Escherichia coli* (D) were scanned against each matrix to find matches to the corresponding DNA sequence motifs. The scanning method assigned a score (maximum 100) to each match site. Scans were performed using a range of different threshold minimum scores. For each threshold, we counted the number of match sites (with a score of equal to or greater than the threshold) found in coding and in non-coding DNA (*i.e.* intragenic and intergenic sites respectively). The ratio of the number of intergenic sites to the number of intragenic sites is plotted for each threshold level that was used. The matrices are further described in Table 1.

Methods

Sequence data were downloaded from the NCBI FTP site [54]. A series of Perl scripts were developed to implement a similar method to that described by Li *et al*. [11] for identification of potential regulatory signals in intergenic regions of complete genome sequences. These scripts are included in additional file 13 [scripts.tar] and can also be freely downloaded from the corresponding author's website [55].

The identification of potential regulatory motifs within a given genome consisted of several steps. First, the script 'get_significant_dyads.pl' (see additional file 13 [scripts.tar]) was used to find statistically over-represented dimers or 'dyads'. Secondly, the significantly over-represented dyads were grouped into clusters of dyads sharing high sequence similarity (using script 'cluster_dyads.pl'). Next, for each cluster, the script 'find_matches.pl' extracted short sequence strings that matched the dyads in that cluster. Then for each cluster, the associated set of sequence matches was aligned using ClustalW [56]. Finally each alignment was converted into a position-specific weight matrix (using script 'make_matrix.pl'), which was used to search the genome sequence for close matches. Each step is described in more detail below.

Identification of significantly over-represented dyads

A dyad is a DNA sequence of the form W_1NxW_2; that is two words, W_1 and W_2, separated by x residues. The pair of conserved oligonucleotides correspond to residues that

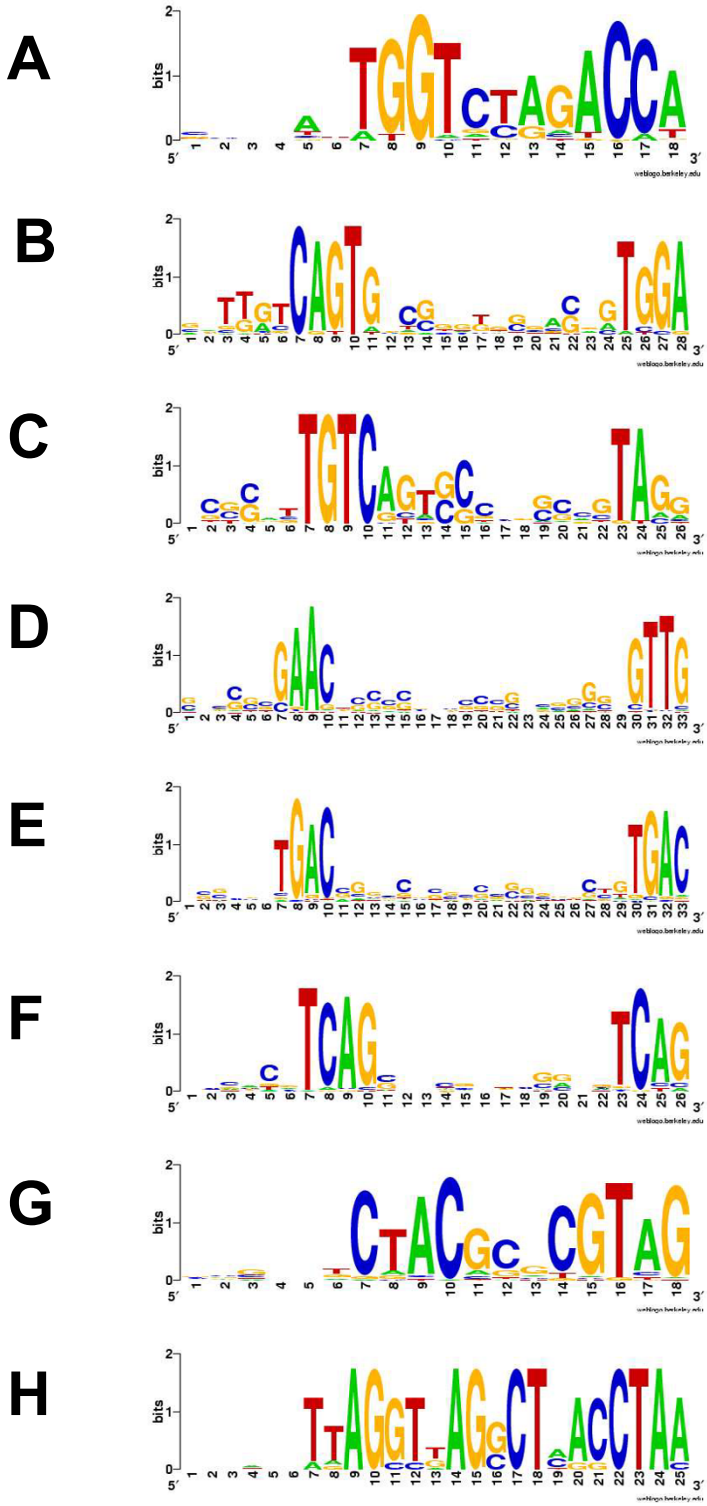

Figure 2
Consensus sequences as indicated from Table 1.

enter into direct contact with the DNA-binding domain of the transcription factor. Their pairing is due to the fact that many bacterial transcription factors form dimers, with each unit binding to a similar small element, accounting for the symmetry of the site [4,11,12].

In these analyses words W_1 and W_2 were each 4 residues long and the spacing, x, was from 4 to 20 residues long. The expected frequency of a word W_1NxW_2 in a given sequence is considered to be the product of the frequencies of the words W_1 and W_2 [10-12]. Given the expected frequency of a dyad and the observed frequency of that dyad within a sequence of known length, it is possible to test whether that dyad is found significantly more frequently than expected. Assuming a Poisson distribution, a P value for the dyad can be calculated as follows. The probability *P* of observing *n(D)* or more occurrences of D is

$$P = \sum_{n \geq n(D)} \frac{[\gamma(D)]^n}{n!} e^{-\gamma(D)}$$

where y(D) is the expected number of occurrences of the dyad D under the null hypothesis that the occurrences of W_1 and W_2 are uncorrelated. The dyad was considered to be significantly over-represented if it had a P value of less than the inverse of the number of dyads considered [10-12]. Thus we would expect less than one of the 'significantly over-represented dyads' to have been included by chance.

In identifying significantly over-represented dyads, we only considered non-coding DNA less than 300 bp upstream of a predicted translational start codon. This upper limit was chosen because it almost all known regulatory sites in bacteria fall within 300 bp of the coding region [59].

Clustering of the significantly over-represented dyads

Many of the over-represented dyads represented different but overlapping versions of the underlying sequence motif [10-12]. To cluster the dimers, a pairwise similarity score was computed for each aligned pair of dyads D_1 and D_2 using the same method as [10-12]. Each dyad was considered as a vertex in a graph object. For each pair of dyads, D_1 and D_2, if the pairwise score was greater than or equal to a threshold of 7, an edge was added to the graph to connect D_1 and D_2. This procedure led to a graph object comprising a forest of a few large trees. Some rather dissimilar dyads were linked to each other via a series of several edges; therefore, the trees needed to be pruned to generate a larger number of compact trees that contained only a few closely related dyads. The process of pruning consisted of removing any edges linking pairs of dyads whose pairwise score was below a threshold of 5.

Generation of weight matrices (PSWMs) and prediction of regulatory sites

For each tree of dyads in the graph object, we compiled a list of all intergenic matches to any dyad within that tree (using script find_matches.pl, see additional file 13 [scripts.tar]). We only included matches in non-coding DNA less than 300 bp upstream of a predicted translational start codon. An ungapped multiple alignment was generated for each list of matches using ClustalW. Each multiple alignment was then converted to a position-specific weight matrix (PSWM) [58] using the make_matrix.pl script (see additional file 13 [scripts.tar]). Genome sequences were searched against PSWMs using the promscan.pl script. The entire genome sequence was scanned such that every window (of the same number of bases as the length of the matrix) was assigned a score using the matrix. This score, known as the Kullback-Leibler distance, reflects the theoretical binding energy of the DNA protein interaction [57] and is calculated using the formula:

$$I_{seq}(i) = \sum_b f_{b,i} \log_2 \frac{f_{b,i}}{p_b}$$

where i is the position within the site, p_b is the frequency of that base in the genome, and $f_{b,i}$ is the observed frequency of each base at that position (from the matrix). To avoid taking the logarithm of zero, a 'pseudocount' was used such that where $f_{b,i}$ was equal to zero, the value of $f_{b,i} \log_2 \frac{f_{b,i}}{p_b}$ was taken as zero. Values for p_b were calculated from the percentage G+C content of the genome sequence. Scores were normalised such that 100 is the highest possible score for a sequence window.

Authors' contributions

DJS conceived of the study, wrote the Perl scripts, carried out the bioinformatic analysis, and wrote the manuscript. SDB and JK also contributed to interpretation of the data and writing the manuscript.

Additional material

Additional File 10
A list of statistically over-represented dyads in S. coelicolor *non-coding DNA.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2180-4-14-S10.txt]

Additional File 11
A list of the 2497 matrices generated from the statistically over-represented dyads in S. coelicolor *non-coding DNA.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2180-4-14-S11.txt]

Additional File 12
The protein classification scheme.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2180-4-14-S12.txt]

Additional File 1
Table 2. Matches to matrix 2083 in the non-coding DNA of S. coelicolor.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2180-4-14-S1.pdf]

Additional File 9
Table 10. Matches to matrix 2083 in the non-coding DNA of S. coelicolor.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2180-4-14-S9.pdf]

Additional File 2
Table 3. Matches to matrix 2083 in the non-coding DNA of S. coelicolor.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2180-4-14-S2.pdf]

Additional File 3
Table 4. Matches to matrix 2083 in the non-coding DNA of S. coelicolor.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2180-4-14-S3.pdf]

Additional File 4
Table 5. Matches to matrix 2083 in the non-coding DNA of S. coelicolor.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2180-4-14-S4.pdf]

Additional File 5
Table 6. Matches to matrix 2083 in the non-coding DNA of S. coelicolor.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2180-4-14-S5.pdf]

Additional File 6
Table 7. Matches to matrix 2083 in the non-coding DNA of S. coelicolor.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2180-4-14-S6.pdf]

Additional File 7
Table 8. Matches to matrix 2083 in the non-coding DNA of S. coelicolor.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2180-4-14-S7.pdf]

Additional File 8
Table 9. Matches to matrix 2083 in the non-coding DNA of S. coelicolor.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2180-4-14-S8.pdf]

Additional File 13
Perl scripts used to perform the analysis.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2180-4-14-S13.tar]

Acknowledgements
DJS is supported by the MRC and is also grateful to Alex Bateman and Sam Griffiths-Jones for useful discussions.

References
1. Hertz GZ, Hartzell GW 3rd, Stormo GD: **Identification of consensus patterns in unaligned DNA sequences known to be functionally related.** *Comput Appl Biosci* 1990, **6:**81-92.
2. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: **Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment.** *Science* 1993, **262:**208-214.
3. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers.** *Proc Int Conf Intell Syst Mol Biol* 1994, **2:**28-36.
4. van Helden J, Rios AF, Collado-Vides J: **Discovering regulatory elements in non-coding sequences by analysis of spaced dyads.** *Nucleic Acids Res* 2000, **28:**1808-1818.
5. Sinha S, Tompa M: **A statistical method for finding transcription factor binding sites.** *Proc Int Conf Intell Syst Mol Biol* 2000, **8:**344-354.
6. Vanet A, Marsan L, Labigne A, Sagot MF: **Inferring regulatory elements from a whole genome. An analysis of *Helicobacter pylori* sigma(80) family of promoter signals.** *J Mol Biol* 2000, **297:**335-53.
7. McCue L, Thompson W, Carmack C, Ryan MP, Liu JS, Derbyshire V, Lawrence CE: **Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes.** *Nucleic Acids Res* 2001, **29:**774-782.

8. Rajewsky N, Socci ND, Zapotocky M, Siggia ED: **The evolution of DNA regulatory regions for proteo-gamma bacteria by interspecies comparisons.** *Genome Res* 2002, **12:**298-308.
9. McGuire AM, Hughes JD, Church GM: **Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes.** *Genome Res* 2000, **10:**744-757.
10. van Helden J, Andre B, Collado-Vides J: **Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies.** *J Mol Biol* 1998, **281:**827-842.
11. Li H, Rhodius V, Gross C, Siggia ED: **Identification of the binding sites of regulatory proteins in bacterial genomes.** *Proc Natl Acad Sci USA* 2002, **99:**11772-11777.
12. Mwangi MM, Siggia ED: **Genome wide identification of regulatory motifs in Bacillus subtilis.** *BMC Bioinformatics* 2003, **4:**18.
13. Chater KF: **Taking a genetic scalpel to the *Streptomyces* colony.** *Microbiol* 1998, **144:**1465-1478.
14. Bentley SD, Chater KF, Cerdeno-Tarraga AM, Challis GL, Thomson NR, James KD, Harris DE, Quail MA, Kieser H, Harper D, Bateman A, Brown S, Chandra G, Chen CW, Collins M, Cronin A, Fraser A, Goble A, Hidalgo J, Hornsby T, Howarth S, Huang CH, Kieser T, Larke L, Murphy L, Oliver K, O'Neil S, Rabbinowitsch E, Rajandream MA, Rutherford K, Rutter S, Seeger K, Saunders D, Sharp S, Squares R, Squares S, Taylor K, Warren T, Wietzorrek A, Woodward J, Barrell BG, Parkhill J, Hopwood DA: **Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2).** *Nature* 2002, **417:**141-147.
15. **Protein Classification Scheme** [http://www.sanger.ac.uk/Projects/S_coelicolor/scheme.html]
16. **EcoCyc: Encyclopedia of *Escherichia coli* Genes and Metabolism** [http://www.ecocyc.org/]
17. Karp PD, Riley M, Saier M, Paulsen IT, Collado-Vides J, Paley SM, Pellegrini-Toole A, Bonavides C, Gama-Castro S: **The EcoCyc Database.** *Nucleic Acids Res* 2002, **30:**56-58.
18. Saito A, Fujii T, Miyashita K: **Distribution and evolution of chitinase genes in Streptomyces species: involvement of gene-duplication and domain-deletion.** *Antonie Van Leeuwenhoek* 2003, **84:**7-15.
19. Delic I, Robbind P, Westpheling J: **Direct repeat sequences are implicated in the regulation of two *Streptomyces* chitinase promoters that are subject to carbon catabolite control.** *Proc Natl Acad Sci USA* 1992, **89:**1885-1889.
20. Miyashita K, Fujii T: **Nucleotide sequence and analysis of a gene (*chiA*) for a chitinase from *Streptomyces lividans* 66.** *Biosci Biotechnol Biochem* 1993, **57:**1691-1698.
21. Fujii T, Miyashita K: **Multiple domain structure in a chitinase gene (*chiC*) of *Streptomyces lividans*.** *J Gen Microbiol* 1993, **139:**677-686.
22. Miyashita K, Fujii T, Saito A: **Induction and repression of a *Streptomyces lividans* chitinase gene promoter in response to various carbon sources.** *Biosci Biotechnol Biochem* 2000, **64:**39-43.
23. Ni X, Westpheling J: **Direct repeat sequences in the *Streptomyces* chitinase-63 promoter direct both glucose and chitin induction.** *Proc Natl Acad Sci USA* 1997, **94:**13116-13121.
24. Saito A, Ishizaka M, Francisco PB Jr, Fujii T, Miyashita K: **Transcriptional co-regulation of five chitinase genes scattered on the *Streptomyces coelicolor* A3(2) chromosome.** *Microbiology* 2000, **146:**2937-2946.
25. Saito A, Miyashita K, Biukovic G, Schrempf H: **Characteristics of a *Streptomyces coelicolor* A3(2) extracellular protein targeting chitin and chitosan.** *Appl Environ Microbiol* 2001, **67:**1268-1273.
26. Saito A, Fujii T, Yoneyama T, Miyashita K: ***glkA* is involved in glucose repression of chitinase production in *Streptomyces lividans*.** *J Bacteriol* 1998, **180:**2911-2914.
27. Angell S, Lewis CG, Buttner MJ, Bibb MJ: **Glucose repression in Streptomyces coelicolor A3(2): a likely regulatory role for glucose kinase.** *Mol Genet* 1994, **244:**135-143.
28. Ingram C, Westpheling J: **The glucose kinase gene of *Streptomyces coelicolor* is not required for glucose repression of the *chi63* promoter.** *J Bacteriol* 1995, **177:**3587-3588.
29. Nguyen J: **The regulatory protein Reg1 of *Streptomyces lividans* binds the promoter region of several genes repressed by glucose.** *FEMS Microbiol Lett* 1999, **175:**51-58.
30. Nguyen J, Francou F, Virolle M-J, Guerineau M: **Amylase and chitinase genes in *Streptomyces lividans* are regulated by *reg1*, a pleiotropic regulatory gene.** *J Bacteriol* 1997, **179:**6383-6390.
31. Schlosser A, Weber A, Schrempf H: **Synthesis of the *Streptomyces lividans* maltodextrin ABC transporter depends on the presence of the regulator MalR.** *FEMS Microbiol Lett* 2001, **196:**77-83.
32. Boel G, Mijakovic I, Maze A, Poncet S, Taha MK, Larribe M, Darbon E, Khemiri A, Galinier A, Deutscher J: **Transcription regulators potentially controlled by HPr kinase/phosphorylase in Gram-negative bacteria.** *J Mol Microbiol Biotechnol* 2003, **5:**206-215.
33. Saier MH Jr, Chauvaux S, Cook GM, Deutscher J, Paulsen IT, Reizer J, Ye JJ: **Catabolite repression and inducer control in Gram-positive bacteria.** *Microbiology* 1996, **142:**217-230.
34. Nothaft H, Parche S, Kamionka A, Titgemeyer F: ***In vivo* analysis of HPr reveals a fructose-specific phosphotransferase system that confers high-affinity uptake in *Streptomyces coelicolor*.** *J Bacteriol* 2003, **185:**929-937.
35. Kamionka A, Parche S, Nothaft H, Siepelmeyer J, Jahreis K, Titgemeyer F: **The phosphotransferase system of *Streptomyces coelicolor*.** *Eur J Biochem* 2002, **269:**2143-50.
36. Parche S, Nothaft H, Kamionka A, Titgemeyer F: **Sugar uptake and utilisation in Streptomyces coelicolor: a PTS view to the genome.** *Antonie Van Leeuwenhoek* 2000, **78:**243-251.
37. Parche S, Schmid R, Titgemeyer F: **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* 1999, **265:**308-317.
38. Wang F, Xiao X, Saito A, Schrempf H: ***Streptomyces olivaceoviridis* possesses a phosphotransferase system that mediates specific, phosphoenolpyruvate-dependent uptake of N-acetylglucosamine.** *Mol Genet Genomics* 2002, **268:**344-351.
39. Ahel I, Vujaklija D, Mikoc A, Gamulin V: **Transcriptional analysis of the *recA* gene in *Streptomyces rimosus*: identification of the new type of promoter.** *FEMS Microbiol Lett* 2002, **209:**133-137.
40. Helmann JD: **The extracytoplasmic function (ECF) sigma factors.** *Adv Microb Physiol* 2002, **46:**47-110.
41. Rawlings ND, Tolle DP, Barrett AJ: **MEROPS: the peptidase database.** *Nucleic Acids Res* 2004, **32:**D160-164.
42. Suzuki M, Taguchi S, Yamada S, Kojima S, Miura KI, Momose H: **A novel member of the subtilisin-like protease family from *Streptomyces albogriseolus*.** *J Bacteriol* 1997, **179:**430-438.
43. Busenlehner LS, Pennella MA, Giedroc DP: **The SmtB/ArsR family of metalloregulatory transcriptional repressors: Structural insights into prokaryotic metal resistance.** *FEMS Microbiol Rev* 2003, **27:**131-143.
44. Kim JS, Kang SO, Lee JK: **The protein complex composed of nickel-binding SrnQ and DNA binding motif-bearing SrnR of *Streptomyces griseus* represses *sodF* transcription in the presence of nickel.** *J Biol Chem* 2003, **278:**18455-18463.
45. Miranda-Rios J, Morera C, Taboada H, Davalos A, Encarnacion S, Mora J, Soberon M: **Expression of thiamin biosynthetic genes (*thiCOGE*) and production of symbiotic terminal oxidase cbb3 in *Rhizobium etli*.** *J Bacteriol* 1997, **179:**6887-6893.
46. Nudler E, Mironov AS: **The riboswitch control of bacterial metabolism.** *Trends Biochem Sci* 2004, **29:**11-17.
47. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR: **Rfam: an RNA family database.** *Nucleic Acids Res* 2003, **31:**439-441.
48. Challis GL, Hopwood DA: **Synergy and contingency as driving forces for the evolution of multiple secondary metabolite production by *Streptomyces* species.** *Proc Natl Acad Sci USA* 2003, **100(Suppl 2):**14555-14561.
49. Ikeda H, Ishikawa J, Hanamoto A, Shinose M, Kikuchi H, Shiba T, Sakaki Y, Hattori M, Omura S: **Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*.** *Nat Biotechnol* 2003, **21:**526-531.
50. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE 3rd, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Krogh A, McLean J, Moule S, Murphy L, Oliver K, Osborne J, Quail MA, Rajandream M-A, Rogers J, Rutter S, Seeger K, Skelton J, Squares S, Squares R, Sulston JE, Taylor K, Whitehead S, Barrell BG: **Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence.** *Nature* 1998, **393:**537-544.
51. Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y: **The complete genome sequence of *Escherichia coli* K-12.** *Science* 1997, **277:**1453-1474.

52. Homerova D, Knirschova R, Kormanec J: **Response regulator ChiR regulates expression of chitinase gene, chiC, in Streptomyces coelicolor.** *Folia Microbiol (Praha)* 2002, **47:**499-505.
53. Stock AM, Robinson VL, Goudreau PN: **Two-component signal transduction.** *Annu Rev Biochem* 2000, **69:**183-215.
54. **NCBI Microbial Genomes** [ftp://ftp.ncbi.nih.gov/genomes/Bacteria/]
55. **Perl Scripts for Prediction of Regulatory Motifs** [http://www.promscan.uklinux.net/dyad_scripts/]
56. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD: **Multiple sequence alignment with the Clustal series of programs.** *Nucleic Acids Res* 2003, **31:**3497-3500.
57. Stormo GD: **DNA binding sites: representation and discovery.** *Bioinformatics* 2000, **16:**16-23.
58. Frech K, Herrmann G, Werner T: **Computer-assisted prediction, classification, and delimitation of protein binding sites in nucleic acids.** *Nucleic Acids Res* 1993, **21:**1655-1664.
59. Gralla JD, Collado-Vides J: **Organization and function of transcription regulatory elements.** In: *Escherichia coli and Salmonella: Cellular and Molecular Biology* 2nd edition. Edited by: Neidhardt FC. Washington, D.C., ASM Press; 1996:1232-1245.