Supplementary Materials:
Extracting regulator activity profiles by integration of de novo motifs and expression data: characterising key regulators of nutrient depletion responses in Streptomyces coelicolor

February 13, 2012

Mudassar Iqbal1,∗, Yvonne Mast2, Rafat Amin2, David Hodgson3, The STREAM consortium, Wolfgang Wohlleben2, Nigel Burroughs4,∗

1Multidisciplinary Centre for Integrative Biology (MyCIB), School of Biosciences, University of Nottingham, Nottingham, UK
2Department of Microbiology/Biotechnology, Faculty of Science, University of Tübingen, Germany.
3School of Life Sciences, University of Warwick, Coventry, U.K.
4Warwick Systems Biology Centre, University of Warwick, Coventry, CV7 4AL.
∗Corresponding authors. N.J.Burroughs@warwick.ac.uk, mudassar.iqbal@nottingham.ac.uk
This document contains an extended methods, a description of the MCMC algorithm, some additional results, and supplementary figures.

Contents

1 Supplementary Methods
  1.1 Experimental Methods .................................................. 2
  1.2 Motif Prediction and Enrichment Analysis ............................ 3
  1.3 A Factor Model: Integrating Motif and Expression Data ............. 5
  1.4 Parameter Inference using a Markov chain Monte Carlo (MCMC) algorithm 
    1.4.1 Convergence Issues: Metropolis Coupled MCMC ................ 9
    1.4.2 Convergence Monitoring .......................................... 10

2 Supplementary Results
  2.1 Gel-shift Analysis for Predicted Motifs ............................. 10
  2.2 Additional Motifs-regulator Identifications ........................ 12
    2.2.1 Motif 27 .......................................................... 12
    2.2.2 Motif 51: Analogues of E. coli regulators Zur/Fur ............. 12

3 Supplementary Figures ........................................................................ 14

1 Supplementary Methods

1.1 Experimental Methods

Time-series culture conditions. Fermentations were carried out as in [1], see Methods in main text.

Culture conditions: Gel-shift Protocol. For EMSA studies S. coelicolor M145 wild-type spores were each inoculated in 200 ml of S-medium (peptone, 4 g/l, yeast extract, 4 g/l, K₂HPO₄, 4 g/l, KH₂PO₄, 2 g/l, glycine, 10 g/l, glucose, 10 g/l, MgSO₄, 0.5 g/l; [2]), glutamate depletion medium (Na-glutamate, 15 g/l; glucose, 40 g/l; MgSO₄, 2.0 mM; phosphate, 9.2 mM; trace element solution [3]; pH 7.0), and phosphate depletion medium (Na-glutamate, 55.2 g/l; glucose, 40 g/l; MgSO₄, 2.0 mM; phosphate, 4.6 mM; supplemented minimal medium trace element solution [3]; pH 7.0) and incubated for 3 days in 500-ml Erlenmeyer flasks with steel springs on an orbital shaker (180 rpm) at 28°C. 50 ml of the preculture were harvested by centrifugation for 10 min at 5000 rpm and were washed two times with the respective main culture medium. Cells were then transferred to 400 ml of the respective main culture medium and incubated in 1-l Erlenmeyer flasks with steel springs for 4 days at 28°C on an orbital shaker (180 rpm). 50 ml culture samples were taken for several time points (8 h, 24 h, 42 h, 48 h, and 72 h). Cells were harvested by centrifugation at 5000 rpm and 4°C for 10 min and disrupted by
using French press (American instruments) with two consecutive passages at 1000 p.s.i. Cell debris were harvested by centrifugation at 10000 rpm and 4°C for 20 min and cell lysates were extracted for EMSA studies.

1.2 Motif Prediction and Enrichment Analysis

To find binding sites throughout the genome, we used a method originally developed in [4] which finds statistically over-represented dyad-type motifs. The motif detection pipe-line consists of the following four steps [5]:

1. Identification of over-represented dyads: The significance of a dyad $D$ is calculated as followed: the expected number of occurrences of dyad $D$ under an independence model is calculated using the total number of occurrences of two words $W_1$ and $W_2$, i.e. $n(W_1)$ and $n(W_2)$ respectively,

$$y(D) = L_{eff}(D) \frac{n(W_1)}{L_{eff}(W_1)} \frac{n(W_2)}{L_{eff}(W_2)}, \quad (1)$$

where, $L_{eff}(M) = \sum_r (L(r) - L(M) + 1)$ is the number of independent positions in the upstream genome sequence data where a motif $M$ of length $L(M)$ can be placed. The summation index $r$ runs over regulatory regions of all genes, each with a length $L(r)$. Given this expected number of occurrences $y(D)$ and assuming occurrences are Poisson distributed (independent), a $P$-value (probability of observing $n(D)$ or more occurrences) can be assigned to dyad $D$ by,

$$P = \sum_{n \geq n(D)} \frac{[y(D)]^n}{n!} e^{-y(D)} \quad (2)$$

A dyad is considered significant if $P < \frac{1}{N_d}$ where $N_d$ is total number of dyads examined. In our experiments we used dyads with word lengths of 4 and 5 (i.e. words $W_1, W_2$ are of lengths 4 or 5) and sequence data comprising up to 300 bases upstream of the predicted translational start codon.

2. Clustering of Significant Dyads: Statistically over-represented dyads found in step 1 might represent different overlapping versions of the same underlying sequence motif; hence they are clustered using the method employed in [5]

3. Building the Motif Models and Predicting Regulatory Sites: For each dyad cluster found in step 2, a list of intergenic sequence matches is compiled. An ungapped multiple alignment is built for each dyad cluster using ClustalW [6]. These alignments are used to build a hidden Markov model HMM for each dyad cluster (a putative motif) using HMMER (http://hmmer.wustl.edu/) software. Then using this model, the entire genome sequence was searched and the corresponding hits listed for each motif model using HMMER software with a default score of 10.
4. Parsing the Output: Genome wide matches found for all the motifs in step 3 are parsed using a user-provided threshold applied to the score.

We further modify this motif level content of genes to the operon level, i.e., we count the hits of motifs in operons rather than individual genes and associate the motif to all genes in the operon. We use the operons defined for \textit{S. coelicolor} by Charaniya \textit{et al.} \cite{7}.

**Enrichment of Motifs in Differentially Expressed Genes**

Applying the above motif search to \textit{S. coelicolor} yielded in excess of 2000 motifs. To restrict to informative motifs, i.e., those with an explanatory capacity for the experimental data, we filtered to those motifs that are significantly enriched in the active sets of genes in our experimental data. Specifically, for each time course experiment we identified the differentially expressed (DE) genes (≤ 1000 genes for each time series) and grouped these genes into clusters with similar expression patterns. We then determined those motifs that are significantly enriched in the sets of DE genes, or the respective clusters, in a two step process as as follows:

1. For a motif \(m\), we denote by \(N_m\) the total number of unique operon hits and by \(K_m\) the number of hits within these \(N_m\) operons which are in the DE set (or a cluster respectively). Under the hyper-geometric model (describing the number of successes without replacement), a P-value can be calculated for each motif, i.e., if \(X_m\) denotes the hits for motif \(m\), then the probability of having more than \(K_m\) counts in DE is:

\[
P_{val}(m) = PR(X_m \geq K_m) = 1 - \sum_{i=0}^{K_m-1} P(X_m = i)
\]  

(3)

where \(P(X_m = i)\) is the probability of \(i\) hits under the hypergeometric model. The enrichment ratio for individual motifs can be calculated with respect to a specific set of genes, i.e., enrichment ratio for motif \(m\) (\(E_m\)) in DE is simply the observed counts of motif \(m\) in DE divided by the expected number of counts in that set.

2. Since we are testing a large number of motifs simultaneously, we need a multiple testing correction. We use a Benjamini and Hochberg correction ([8]) for controlling the false discovery rate as follows: rank the list of P-values corresponding to the \(T\) motifs (or in other words the \(T\) hypotheses that we are testing), i.e., order \(P_1 \leq P_2 \leq \ldots \leq P_T\), with \(P_i\) the P-value of null hypothesis \(H_i\). For a given q-value \(q^*\), let \(k\) be the largest \(i\) for which \(P_i \leq \frac{i}{T} q^*\). Then reject all null hypotheses \(H_i\), \(i = 1, 2, \ldots, k\). For the DE set of genes on TS1, Fig. S-1 shows how the \(p\)-values and \(q\)-values are related.
Comparison of P-values and Q-values

Figure S-1: P-value (left panel) and q-value (right panel) for motifs enriched in DE genes of TS1. This only includes motifs with q-values not more than 10%. Here \( q_i = \frac{p_i}{T} \) where \( T \) is the total number of motifs or tests and \( i \) is the rank of ordered p-values \( p_i \).

1.3 A Factor Model: Integrating Motif and Expression Data

We used a factor model to describe the constrained relationship between the expression data of \( N \) genes across \( M \) time points and the \( L \) binding motifs. In matrix form this is given by

\[
E = AP + \Gamma \tag{4}
\]

where (we use lower-case for elements of matrices) \( e_{it} \) is the expression of gene \( i \), \( p_{jt} \) the (unknown) activity of TF \( j \) both at time \( t \), \( a_{ij} \) the control strength of the transcriptional regulator through motif \( j \) on gene \( i \), and \( \gamma_{it} \) Gaussian noise, assumed homogeneous in time. Connectivity is restricted by the requirement for the presence of the motif in a gene’s upstream region, i.e. \( a_{ij} \) is zero if there is no motif \( j \) in the upstream regions of gene \( i \). To allow for false positives in the motif search, or inactive binding sites, we parametrise the active regulatory network connectivity by \( Z \), i.e. \( z_{ij} = 1 \) if TF \( j \) regulates gene \( i \). Thus, \( a_{ij} = 0 \) if \( z_{ij} = 0 \), and \( z_{ij} \) can only possibly be 1 if there is a binding site for motif \( j \) upstream of gene \( i \) (or the operon containing gene \( i \)). This describes a two-layered bipartite regulatory network, Fig S-2, where edges are between TF’s and their target genes (corresponding to \( z_{ij} = 1 \)).
1.4 Parameter Inference using a Markov chain Monte Carlo (MCMC) algorithm

We use a Bayesian methodology for the inference of the model parameters, i.e. $A$, $P$, noise variance $\sigma_i^2$ as well as the network structure/connectivity (the $Z$ matrix). The prior on the network structure is provided by the binding site data ($z_{ij}$ can only be non zero if there is a binding site for motif $j$ in the operon containing gene $i$) while for the rest of the parameters a conjugate prior (\cite{[9]}) is appropriate. This model has an identifiability issue; however provided the constraints due to the presence of the binding motifs impose sufficient sparseness on the potential network the remaining symmetry is highly restricted and, importantly, fixes the identity of the motifs explicitly through the effects on their targets, i.e. there is no interchangeability symmetry, \cite{[9, 10]}. We found no evidence that this was not true for our data/selected motifs.

We choose weak priors apart from the network topology (which is highly constrained by the binding site data). These are given as follows,

- $P$: Each $p_{jt}$ is considered a priori independent with distribution $N(0, \sigma_p^2)$.
- $A$: The connectivity matrix $A$ is constrained by the network topology, i.e., $a_{ij} = 0$ if $z_{ij} = 0$. A Gaussian priori is imposed on the non-zero entries, i.e. $a_{ij} \mid z_{ij}=1 \sim N(0, \sigma_a^2)$. We follow the recommendations made in the earlier work \cite{[9]} and use an informative prior on $P$ i.e., $\sigma_p = 1$ and uninformative prior for $A$ ($\sigma_a = 100$). This removes the rescaling symmetry (non-identifiability) since only products $a_{ij}p_{jt}$ are meaningful.
• Noise Precision $\sigma^{-2}_i$: We use a Gamma distribution prior, $\Gamma(\alpha_i, \beta_i)$ with parameters values appropriate for a weak prior, $\alpha = 0.9, \beta = 0.01$. This is much broader than the posterior on our data, Fig. S-4, confirming it is weak.

The posterior distribution of the model parameters is sampled using an MCMC methodology. We exploit the conditional independence structure among our parameters by using a collapsed Gibbs sampler. The following conventions are used, [9]:

• $x^y = \prod_{i=1}^{r} x_i^{y_i}$
• $a[z]$ is the vector of non-zero elements of $a$ corresponding to $z_{ij} = 1$ in the current state.
• Let $P$ be a vector with the number of rows equal to the number of elements of vector $z$. Then $P[z]$ is the sub-vector obtained by selecting the rows of $P$ corresponding to non-zero elements of $z$.
• If the matrix $A$ has same dimension as $Z$, then $A_Z$ is identical to original matrix $A$ except that all its elements corresponding to zeros in matrix $Z$ are also zero, i.e. $(a_Z)_{ij} = a_{ij}z_{ij}$.

The conditional distributions of the model parameters are given by,

$$P(z^i|P, \sigma^2) \propto \frac{\pi^{(z^i)}(1 - \pi^{(1-z^i)})}{\sigma^{|z^i|}_a} \cdot \text{det}(P[z^i]P[z^i]'/\sigma^2_i + I_{|z^i|}/\sigma^2_z)^{-1/2} \cdot \exp \left\{ \frac{1}{2\sigma^2_i} e^i P[z^i]' \left( \frac{P[z^i]P[z^i]'}{\sigma^2_i} + \frac{I_{|z^i|}}{\sigma^2_z} \right)^{-1} P[z^i]e^i \right\}$$

$$a_i|(P, Z, \sigma^2) \sim N \left( \Sigma_a, P[Z_i]e^i/\sigma^2_i, \Sigma_a \right)$$

$$p_t|(A, Z, \sigma^2) \sim N \left( \Sigma_{pt}, A_z^*\Sigma^{-1}e_t, \Sigma_{pt} \right)$$

$$1/\sigma^2_i(A, Z, P) \sim \text{Gamma} \left( \tilde{\alpha}_i, \tilde{\beta}_i \right)$$

where we define the following matrices: $I$ is the identity matrix and $\Sigma_{ai} = \left( \frac{P[z^i]P[z^i]'}{\sigma^2_i} + \frac{I_{|z^i|}}{\sigma^2_z} \right)$, $\Sigma_{pt} = \left( A_z^*\Sigma^{-1}A_z + \frac{I_z}{\sigma^2_z} \right)^{-1}$ and parameters $\tilde{\alpha}_i = \alpha_i + M/2$. The Gamma distribution parameters are $\tilde{\beta}_i = \beta_i + \sum_{t=1}^{M}(e_{it} - \sum_{j=1}^{L}a_{ij}p_{jt})^2/2$. In the conditional for $Z$ in equation 5 variable $A$ is integrated out (the collapsing step); the resulting conditional distribution for $Z$ is independent of $A$.

M-H step within Gibbs for $Z$-Update

The original work [9] used a Gibbs update for the topology $Z$ which required a computationally expensive normalisation across all possible topologies. We found this to be
unworkable on our data. We therefore imposed a local M-H move $Z$ by restricting moves to within a certain Hamming distance, typically up to 3 changes. Given the current $Z$ state, we choose surfaces of Hamming distances 1, 2 and 3 around the current state and a new state $Z'$ is selected from the associated ball, Fig. S-3. Specifically, the proposed state $Z'$ is drawn from the normalised conditional distribution $p(Z'|P,\sigma^2)$ of equation 5 restricted to the ball around the current state, $B(Z)$, of Hamming distance $d$. The normalization constant for the $B(Z)$ is given by,

$$N_Z = \sum_{Z' \in B(Z)} P(Z')$$

(6)

where $P(Z')$ denotes the right hand side of equation 5.

![Figure S-3: Hamming distance surfaces of distances $d = 1, 2, 3$ around the current and new states $Z, Z'$. The number of states are $n$, $\frac{n(n-1)}{2}$ and $\frac{(n-1)(n-2)}{2}$ respectively, where $n$ is the number of binding sites found by the motif search (allowing for operons).](image)

Now propose a $Z'$ with probability $P(Z')/N_Z$. Now we place a similar ball $B(Z')$ of radius $d$ around the new location $Z'$ and calculate the corresponding conditional probability within the ball of all possible states in $B(Z')$. The corresponding normalization constant is given by

$$N_{Z'} = \sum_{Z'' \in B(Z')} P(Z'')$$

(7)

Finally, the acceptance probability of the new state $Z'$ is given by the MH ratio:

$$\alpha_{Z \rightarrow Z'} = \min \left( 1, \frac{P(Z') Q(Z|Z')}{{P(Z) Q(Z'|Z)}}, \frac{N_Z}{N_{Z'}} \right)$$

(8)
1.4.1 Convergence Issues: Metropolis Coupled MCMC

The MC only samples the desired posterior distribution if it has converged [11]. However, as is often the case with high dimensional MCMC algorithms that explore a network topology, the network variables \((Z)\) have poorer mixing than the rest of the variables in the MC, i.e., the burn-in and sampling times may be unrealistically long for ‘large’ networks. This problem is largely caused by the presence of a few motifs with large numbers of hits in the genome (possibly a result of the high GC content of \(S.\ coelicolor\) and the presence of a high fraction of false positives in the binding sites data). To deal with these issues, we implemented a MCMC methodology that handles this type of mixing problem, specifically Metropolis coupled MCMC, originally proposed in [11, 12]. MCMCMC has been used in other biological inference problems, e.g. [13]. This method runs multiple parallel Markov chains, some of which are ‘heated’ (i.e., the likelihood is raised to some power) which flattens the landscape and hence the chain has a reduced tendency to get trapped in local minima as compared to normal (‘unheated’ or ‘cold’) chain. These chains are allowed to attempt a swap of states periodically during the runs, and in this way the cold chain, through successive exchange with hotter chains, explores with greater efficiency the distribution resulting in improved convergence and mixing properties.

Similar to [13], we use \(M\) Markov chains with associated temperatures \(\theta_i\) for chain \(i\), where \(i\) labels the chains, \(i = 1,..M\), in an incremental fashion so that the first chain \((i = 1)\) is always the cold chain from which the samples of the posterior are drawn. The temperature for chain \(i\) is \(\theta_i = 1/[1 + \Delta T \cdot (i - 1)]\) where the parameter \(\Delta T > 1\) is chosen to give a swap acceptance rate between 20% to 60%.

For chain \(i\) at temperature \(\theta_i\) we raise the likelihood to the power \(\theta_i\). Given the form of our conditional posteriors in equations 5, this task amounts to a transformation of \(\sigma_i\) to \(\sigma_i/\theta\) in the conditionals for the parameters \(Z, A, P\) while for \(\sigma^2_i\), the conditional distribution parameters \(\tilde{\alpha}\) and \(\tilde{\beta}\) are altered to,

\[
\tilde{\alpha}_i = \alpha_i + \frac{M\theta}{2} \tag{9}
\]

\[
\tilde{\beta}_i = \beta_i + \theta \cdot \sum_{t=1}^{M} \left( \epsilon_{it} - \sum_{j=1}^{L} a_{ij}p_{jt} \right) \tag{10}
\]

The rate of chain exchange must be correctly balanced so that the cold chain converges to the desired posterior. This is achieved as follows, [13]. Define the full posterior \(P(\psi|E) \propto L(\psi|E) \cdot \pi(\psi)\) where \(\psi = (Z, A, P, \sigma^2)\) is the parameter set in our model and \(L(\psi|E), \pi(\psi)\) are the likelihood on data \(E\) and prior respectively. For the heated chain the corresponding posterior is \(\propto [L(\psi|E)]^{\theta_i} \cdot \pi(\psi)\). Then the acceptance rate for exchanging the states \(\psi_i, \psi_j\) of the chains \(i, j\), temperatures \(\theta_i, \theta_j\) is given by the MH ratio,

\[
\alpha_{\text{swap}} = \min \left( 1, \frac{L(\psi_j|E)^{\theta_i} \cdot L(\psi_i|E)^{\theta_j}}{L(\psi_i|E)^{\theta_i} \cdot L(\psi_j|E)^{\theta_j}} \right) \tag{11}
\]

since the priors cancel.
1.4.2 Convergence Monitoring

We employed multiple chains to confirm convergence. We found that when the topology variables converged, typically all the other variables had done so for both the MCMC and MCMCMC algorithms. A typical run is shown in Fig. S-4. On our data we used a burn-in of 100 iterations and employed subsampling after every 5th iteration. Simulations were run until the convergence was reached.

Figure S-4: Convergence diagnostics: (a) Likelihood (sum of squared error) for two different chains (blue and red), (b) Mean values of MCMC samples for Z parameter for two different chains, (c) Mean values of MCMC samples for the noise variance parameter $\sigma^2$ for two different chains, (d) Prior (blue) and posterior (red) distribution comparison for the noise precision parameter $1/\sigma^2$ under the model with a single noise parameter. Data shown in (a) show full run including the burn-ins while in (b-d) only the sampling phase data are used.

2 Supplementary Results

2.1 Gel-shift Analysis for Predicted Motifs

To investigate if the predicted motifs are binding to a regulatory effector, DNA fragments containing the predicted binding motif sequences were tested in electromobility shifts assays (EMSAs). Gel-shifts can indicate only that there is a protein binding these sites under the given conditions, but not whether the actual predicted motif is responsible. As such, it is therefore only a preliminary indication that there is a biological significance
to the predictions and more sophisticated (and time consuming) methods are needed to prove a direct correspondence.

For the EMSA analyses the *S. coelicolor* M145 wild-type strain was grown in different culture media (complex S-medium, glutamate depletion medium, and phosphate depletion medium) and samples were taken for different culture time points (8 h, 24 h, 42 h, 48 h, and 72 h) corresponding to the transcriptomic data sets (see above). The different *S. coelicolor* cell lysates were analysed in standard EMSAs. We designed primers for the different DNA fragments to include only the respective motif but none of the other predicted motif sequences.

We selected 26 of the strongest predictions for testing by gel-sift analysis. We also tested three negative control sequences that did not include any of the motifs. We found evidence of protein binding in at least one of the culture conditions for 5 of the motifs; there was no such shift for any of the control sequences contrSCO4556, contrSCO4752, contrSCO6189 with any of the cell lysates (data not shown). A retarded band was observed (Fig. S-5) upon incubation of SCO3320 with *S. coelicolor* cell lysates from all media (time point 48 h, 72 h), SCO3945 with *S. coelicolor* cell lysates from glutamate and phosphate depleted media (time point 48 h, 72 h), and for SCO4562 with all *S. coelicolor* cell lysates from all time points, SCO6551 with *S. coelicolor* cell lysates from S-medium (time point 48 h). We also found evidence for binding for SCO0079 with *S. coelicolor* M145 cell lysate from S-medium (48 h); SCO0079 was a prediction under an earlier analysis for binding Motif 6 but was removed later as the enrichment filtering step was developed, specifically when it was expanded to include an analysis of the clusters. The best shifts were obtained when the culture was 48 h to 72 h old. The specificity of the binding was verified by a 500-fold excess of unlabelled competitor DNA.

![Figure S-5: EMSAs with DNA upstream regions containing *S. coelicolor* M145 cell lysate from S-medium (72 h) and SCO3320, *S. coelicolor* M145 cell lysate from glutamate depletion medium (72 h) and SCO3945, *S. coelicolor* M145 cell lysate from S-medium (48 h) and SCO4562 cell lysate from S-medium (48 h) and SCO6551. Predictions for motif 6 include SCO3320, SCO3945, SCO4562, and for motif 20, SCO6551. Lanes: 1, control, 2 ng of Cy5 labelled DNA; 2, DNA and 7 ul *S. coelicolor* M145 cell lysate; 3, control, DNA, cell lysate and 500-fold excess of specific DNA (non-labelled); 4, control, cell lysate and 500-fold excess of unspecific DNA (non-labelled).](image)


2.2 Additional Motifs-regulator Identifications

We discuss here two additional motifs with homologs in *E. coli*.

2.2.1 Motif 27

Motif 27, (in cluster 2) shows decreasing activity in all time series, but also substantial inhibition at the time of glutamate depletion, TS5. This specific behaviour in TS5 suggests that it is related to respiration collapse and/or nitrogen starvation [14]. A motif sequence comparison using STAMP identifies *phoB* in *E. coli* as the best hit for motif 27, \( p = 5.2 \times 10^{-4} \). However, the orthologue of *E. coli* PhoB is PhoP, [15] for which we have assigned motif 22. This hit is thus clearly coincidental. The next best hits are *rpoD18*, *tyrR*, and *rpoD15*, but these are not sufficiently strong hits to be meaningful.

2.2.2 Motif 51: Analogues of *E. coli* regulators Zur/Fur

Motif 51 is one of the few motifs in our analysis which shows evidence of being switched off in all the experimental conditions, *i.e.*, all potential links have a posterior probability \( P(z_{ij} = 1) < p_j \). However, we did find a significant match for the motif to the Zur/Fur family of regulators. These are heavy metal sensitive genes in *E. coli*, Zur being a zinc-specific repressor [16]. Supplementary Fig. S-6 we show the sequence signatures of motif 51, the matching sequence logo of the regulator Zur (STAMP p-value = 5.9e^{-07}) as well as the predicted activity profile of motif 51. The lack of a significant signal from its targets is consistent with the lack of zinc starvation.
Figure S-6: Motif 51 (Zur): (a) Sequence logo for motif 51. (b) Zur/Fur logo matching motif 51 (STAMP). (c) Predicted activity profile (red). There are no significant targets. Vertical solid lines (brown) separate three time series, TS1, TS3, and TS5 respectively while the vertical dashed lines (magenta) correspond to the nutrient depletion time in each time series.
Figure S-7: Number of (operon-level) hits in the genome versus GC content of the motifs. GC content of each motif was calculated using all the exact dyad-matching sequences used to build the motif model. Motifs marked in red (55) were used in the analysis reported in the text, having a GC content less than 75% and a number of hits below 300. Only motifs that are determined to be informative in the transcription data (e.g., enriched in DE, see text) are shown.
Figure S-8: Distribution of the number of motifs per gene. Left. Motif counts for the 551 genes having at least one binding site for the 55 motifs. Right. Posterior probability distribution computed from the factor model for the number of significant motifs per gene amongst the 55 motifs and 551 genes in (left).

Figure S-9: Distribution of number of motifs per gene for all 1620 DE genes (55 motifs) with a Poisson fit shown (red) with mean 0.525.
Figure S-10: Posterior for the motif connectivity $\rho$ calculated from the MCMC sampler. Sorted mean values are shown for each motif (triangles), error bars correspond to 25%, 75% percentiles of the posterior distribution. Prior mean value (quantiles 25% and 75%) on $\rho_j$ is shown as dotted horizontal line (black, respectively blue).

Figure S-11: Hierarchical clustering tree for the motifs based on their activity profiles (MATLAB implementation of the hierarchical linkage clustering method).
Figure S-12: Clustered motif activity profiles for the analysis with 72 motifs discussed in the robustness analysis of main text. The motif activity profiles grouped into 9 activity clusters. Individual motif patterns are distinguished by colour. In each of the plots, vertical lines separate the three times series, TS1, TS3, and TS5 respectively.
Figure S-13: Motif 22 (PhoP). Predicted Activity profile (red) along with the mean expression values (blue) and range (shaded) of top five target genes. Vertical solid lines (brown) separate three times series, TS1, TS3, and TS5 respectively while the vertical dashed lines (magenta) correspond to the nutrient depletion time in each time series.
Figure S-14: Activity profiles of Motifs 1 to 30. Shown in red is the (mean) predicted activity of the motif using the factor model while in blue is the average expression profile of the (at most) top five significant targets. The shaded area (light blue) shows the range of the gene profiles of these top significant targets. Vertical lines separate the three time series, TS1, TS3 and TS5 respectively.
Figure S-15: Activity profiles of Motifs 31 to 55. Shown in red is the (mean) predicted activity of the motif using the factor model while in blue is the average expression profile of the (at most) top five significant targets. The shaded area (light blue) shows the range of the gene profiles of these top significant targets. Vertical lines separate the three time series, TS1, TS3 and TS5 respectively.
Figure S-17: Motif Sequence logo in activity cluster 1

Figure S-16: Posterior distribution of $Z$ parameter (a) the distribution over all pooled targets, motifs for $z_{ij}$ while (b), motif 1, (c) motif 15, and (d) motif 51 show three selected motifs in which coloured in red are the values of the posterior probability $z_{ij}$ which are higher than their corresponding prior values (i.e greater than posterior mean of $\rho_j$), and thus determined as significant.
Figure S-18: Motif sequence logos in activity cluster 2

Figure S-19: Motif sequence logos in activity cluster 3
Figure S-20: Motif sequence logos in activity cluster 4

Figure S-21: Motif sequence logos in activity cluster 5

Figure S-22: Motif sequence logo in activity cluster 6. A palindrome
Figure S-23: Motif sequence logos in activity cluster 7. Motif 22 is a PHO box, with a missing letter $G$ at the start of the first dyad.

Figure S-24: Motif sequence logo in activity cluster 8.

Figure S-25: Motif sequence logo in activity cluster 9.
Figure S-26: Motif sequence logos in activity cluster 10

Figure S-27: Binding sites locations for motifs in activity cluster 2. Only those genes (55) which have at least two motif binding sites are shown.
Figure S-28: Distribution of the binding site locations with respect to translation start site for all genes and all motifs in the data set. Left panel shows the distribution of all binding site locations while the right panel shows only those which were significant, i.e. supported by the expression data.

**Supplementary Tables as Excel Documents**

Following supplementary tables are included with this draft as a separate excel file:

- **Table1-AllMotifsTargets**: A complete list of all motifs and their targets (with corresponding probabilities for \( z_{ij} = 1 \)).

- **Table2-MotifTargetAnnotation**: Motifs and their target genes along with the prior (\( \rho_j \)) and posterior probabilities of a motif-gene link. Gene annotation of the target genes is also included.

- **Table3-RegulatorsComparisonConsolidated**: Includes for each predicted motif(regulator), top five matches among the list of known regulators along with their similarity score and annotation.

- **Table4-MotifLocations**: A matrix containing locations of binding sites with respect to translation start site for all genes and all motifs in the data set.

- **Table5-ClusterTargetCorrelations**: Contains tables in one excel sheet (one for each motif activity cluster) for the target overlaps between the motifs in each activity cluster. Each table entry corresponds to the proportion of common targets (defined
as those having posterior $P(z_{ij} = 1 > \rho_j)$ among the motifs $k$ (row) and $j$ (column) relative the number of targets of motif $k$.

- Table6-Cluster7MotifMEME: Lists of the targets of activity cluster 7 (PhoP related) along with their PHO box matching sequences, found using MEME.

- Table7-GelShifsAnalysis: Two tables, containing list of motifs and their target genes tested for experimental verifications, and the list of primers used for EMSA’s.

- Table8-ScoRegulatorsAnnotation: list of the possible DNA binding regulators in $S. coelicolor$ with annotation.

References

[1] Nieselt K. et al (2010) The dynamic architecture of the metabolic switch in *Streptomyces coelicolor*, *BMC Genomics* **11**(10).

[2] Okanishi M., Suzuki K., and Umezawa H. (1974) Formation and reversion of streptomyces protoplasts: cultural condition and morphological study. *J. Gen Micro* **80**, 389-400.

[3] Claessen D., Rink R., de Jong W. et al (2003) A novel class of secreted hydrophobic proteins is involved in aerial hyphae formation in *Streptomyces coelicolor* by forming amyloid-like fibrils. *Genes Dev.* **17**, 1714-26.

[4] Li H., Rhodius V., Gross C. et al (2002) Identification of the binding sites of regulatory proteins in bacterial genomes, *Proc. Natl. Acad. Sci. USA* **99**(18), 11772-7.

[5] Studholme D. J., Bentley S. D., and Kormanec J. (2004) Bioinformatic identification of novel regulatory DNA sequence motifs in *Streptomyces coelicolor*, *BMC Microbiology* **4**(14).

[6] Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignments through sequence weighting, position specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673-4680.

[7] Charaniya S., Mehra S., Lian W. et al (2007) Transcriptome dynamics-based operon prediction and verification in *Streptomyces coelicolor*, *Nucleic Acids Research* **35**(21), 7222-7236.

[8] Benjamini Y. and Hochberg Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing, *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1), 289-300.

[9] Sabatti C. and James G. M. (2006) Bayesian sparse hidden components analysis for transcription regulation networks, *Bioinformatics* **22**(6), 739-746.
Liao J., Boscolo R., Yang Y. *et al* (2003) Network component analysis: Reconstruction of regulatory signals in biological systems. *PNAS*, 100, 15522-15527.

Gilks W. R. and Roberts G. O. (1996) Strategies for improving MCMC. In Gilks, W.R., Richardson, S. and Spiegelhalter (eds), *Markov chain Monte Carlo in Practice*. Chapman and Hall, London, 89-114.

Geyer C. J. (1991) Markov chain Monte Carlo maximum likelihood. In Keramidas (ed.), *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*. Interface Foundation, Fairfax Station, 156-163.

Altekar G., Dwarkadas S., Huelsenbeck J. P. *et al* (2004) Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference, *Bioinformatics* 20(3), 407-415.

Waldvogel E., Herbig A., Battke F. *et al* (2011). The PII protein GlnK is a pleiotropic regulator for morphological differentiation and secondary metabolism in *Streptomyces coelicolor*. *Appl. Microbiol. Biotechnol.* 92, 1219-1236.

Rodríguez-García A., Sola-Landa A., Apel K. *et al* (2009) Phosphate control over nitrogen metabolism in *Streptomyces coelicolor*: direct and indirect negative control of glnR, glnA, glnII and amtB expression by the response regulator PhoP, *Nucleic Acids Res.*, 37(10), 3230-3242.

Patzer S. I. and Hantke K. (2000) The zinc-responsive regulator Zur and its control of the znu gene cluster encoding the ZnuABC zinc uptake system in *Escherichia coli*. *J Biol Chem*, 275(32), 24321-32.