Charles University in Prague
Faculty of Mathematics and Physics

DOCTORAL THESIS

Eva Straková

Identification and modeling of gene expression regulatory networks during streptomycetes germination

Institute of Physics of Charles University

Supervisor of the doctoral thesis: Ing. Jiří Vohradský, CSc.
Study programme: Physics
Specialization: Biophysics, Chemical and Macromolecular Physics

Prague 2013

Acknowledgements

I would like to express my gratitude to my supervisor Ing. Jiří Vohradský, CSc. for inspiring leading of my project, encouragements and advices. My special thanks are due to Bc. Alice Ziková for her tenacity in performing demanding experiments. I would like to thank RNDr. Jan Bobek, Ph.D. for introducing me to the microbiology of *Streptomyces* and for inspiring discussions and comments to the project.

I wish to thank all other collaborators participating in the project and colleagues from the department of Cell and Molecular Microbiology at the Institute of Microbiology (Czech Academy of Sciences), where the project was conducted, for help and friendly working environment.

My thanks also belong to my family and boyfriend for continuous support.

This work was mainly supported through grants of the Czech Science Foundation P302-11-0229 and 310-07-1009, and grant of the Grant Agency of the Charles University under contract No. 17409.

I declare that I carried out this doctoral thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

In Prague, July 23, 2013 Eva Straková

Název práce: Identifikace a modelování regulačních sítí genové exprese v průběhu germinace streptomycet

Autor: Eva Straková

Ústav: Fyzikální ústav Univerzity Karlovy

Školící pracoviště: Microbiologický ústav Akademie věd České republiky

Vedoucí doktorské práce: Ing. Jiří Vohradský, CSc., Microbiologický ústav Akademie věd České republiky

Abstrakt: Streptomycety jsou studovány především kvůli své schopnosti produkovat antibiotika, ale také jako modelový bakteriální organismus s komplexním buněčným cyklem. V této práci byl systémově studován proces klíčení (germinace) *Streptomyces coelicolor* za použití transkriptomických a proteomických metod. Klíčení představuje základní vývojový přechod z klidového období buňky do vegetativní fáze růstu. Během počátečních 5,5 h klíčení byly ve 13 časových bodech měřeny změny exprese mRNA a celková vnitrobuněčná koncentrace proteinů, včetně monitorování aktuání proteosyntézy. Pro kvantifikaci transkriptu byla použita metoda DNA mikročipů, proteiny byly měřeny pomocí dvourozměrné gelové elektroforézy. Modelováním genové expresse byly rekonstruovány genetické sítě a identifikovány funkční skupiny genů regulované určitými sigma faktory. Výstupem modelování byl soubor parametrů, který umožnil simulovat kinetiku regulace mezi sigma faktory a regulovanými geny. Bylo zjištěno, že klíčovou roli během procesu hrají sigma faktory SigR a HrdD, jejichž regulony byly identifikovány. Z transkriptomických i proteomických dat vyplývá, že nejintenzivnější odezva buňky na vnější prostředí probíhá během první hodiny germinace, aktivací různých regulačních mechanismů. Pro porovnání globálních trendů v expresi genů/proteinů byla aplikována metoda analýzy hlavních komponent, která odhalila základní funkční skupiny genů a proteinů spojené s regulačními procesy ovlivňujícími germinaci.

Klíčová slova: germinace, regulační sítě, sigma faktory, *Streptomyces*

Title: Identification and modeling of gene expression regulatory networks during streptomycetes germination

Author: Eva Straková

Department: Institute of Physics of Charles University

Affiliation: Institute of Microbiology Academy of Sciences of the Czech Republic

Supervisor: Ing. Jiří Vohradský, CSc., Institute of Microbiology Academy of Sciences of the Czech Republic

Abstract: Streptomycetes have been studied mostly as producers of antibiotics and for fundamentals of complex bacterial cell development. Here, transcriptomic and proteomic approaches were applied to systems study of *Streptomyces coelicolor* germination as a developmental transition from dormancy to the vegetative stage. The time dynamics of the gene expression levels represented by mRNA and intracellular protein accumulation and synthesis were measured throughout 5.5 h of germination at 13 time points by employing both DNA microarray and two-dimensional gel electrophoresis techniques. Using a numerical model of gene expression, genetic networks were reconstructed and functional groups of genes controlled by the sigma factors were identified. Modeling of the regulatory interactions provided a set of parameters allowing simulate kinetics of gene expression control among the sigma factors and their target genes. Particularly regulons of two sigma factors, SigR and HrdD, were identified. The analysis assigned their key role during the germination process. Analysis of global trends in the gene/protein expression revealed that the full capability of regulatory mechanisms responding to the environmental cues is reached within the first hour of germination, and identified the basic gene/protein functional groups and associated regulatory processes controlling germination.

Keywords: gene regulatory networks, germination, sigma factors, *Streptomyces*

Table of Contents

1 Introduction

1.1 Modeling of gene regulatory networks

From the biological perspective every process in the living cell is under tight control encoded in the genome. General term gene regulation includes a number of sequential processes (the most well comprehended and described are transcription and translation) used to adjust the synthesis of specific gene products to actual requirements of the cell. Thus, in principle, single gene influences expression of other genes and the interacting system creates a network - the gene regulatory network (GRN). According to the results of biological observations, the typical characteristic of GRN is that small constant number of regulatory genes control expression of great set of other genes, called target genes. The complexity of regulatory procedures is simplified for the purposes of GRN modeling to be able to capture essential principles of the particular process. On transcriptional level, the regulation is mostly realized by proteins, specific transcriptional regulators/transcription factors (TF), which are products of special regulatory genes. The TF control transcription of their target genes through DNA-binding associations to promoter sequence regions and thus mediate the transcription initiation.

The computational approach defines inference of GRN as a process of identification of gene interactions from experimental data. There are two broad categories of experimental methodology and relevant reverse-engineering algorithms.

In the first method, the physical binding interactions of TF to regulatory site of target gene promoters are identified. A common experimental approach to this problem is identification of TF's DNA-binding sites by employing genome-wide chromatin immunoprecipitaion (ChIP) experiments (ChIP-chip and ChIP-Seq). TF-promoter interaction are identified also computationally; TF binding sites are predicted by scoring the match of the sequences of given promoter with the given TF's motif. A number of scoring models supported by different mathematical background has been presented. The advance in the field is reviewed in (Hardison and Taylor 2012), however, it is not further discussed in this thesis.

The second method is based on the dynamic nature of transcription and relates the expression level of TF to the expression levels of the other genes controlled by the TFs. This approach was enabled by the possibility to measure the amount of expressed transcript (mRNA molecules) on large-scale level and thus gain the information about transcription activity as a function of time, simultaneously for a large number of genes or even for whole genome. Gene chips used for monitoring relative quantity of mRNA are called microarrays (experimental technique described in paragraph 1.2.1). To observe kinetic changes in mRNA levels as a response to perturbation or examined developmental process, the temporal expression data are measured using multiple microarray experiments. Such time courses can provide useful data source for computational models associated with inferring GRNs.

It should be emphasized that the "dry" computational methods always work in probabilistic nature and the results should be understood in probabilistic terms as well. Thus the bioinformatics identifications cannot achieve a deterministic accuracy of "wet" experimental methods. Importance of the computational methods lies in their low-cost and high-speed in comparison with requirements for the "wet" experiments. Computational approaches are essential in reduction of immense number of combinations of regulatory interactions by identifying only key genes, molecules or pathways for "wet" lab verifications. Additionally, performing virtual *in silico* simulations based on the computational models enable to investigate the system under novel experimental conditions without costly and laborious laboratory techniques.

In summary, to obtain comprehensive picture of interactions governing the cell life it is useful to integrate both biological experiments and computational modeling in a complementary manner. Next chapters will focus on a brief description of basic classes of modeling formalism that have been used for GRNs identification from gene expression time series. More attention is paid to the model based on neural networks (chapter 1.1.7) which was used for the data analysis in this thesis.

1.1.1 Boolean networks

The simplest approximation of the biological systems properties represents Boolean networks. They are direct graphs (Y, E) , where $y_i \in Y,\ 1 < i \leq n$ are nodes described by Boolean variables (Figure 1) and $e_j \in E,\ 1 < j \leq m$ is a set of edges of the graph.

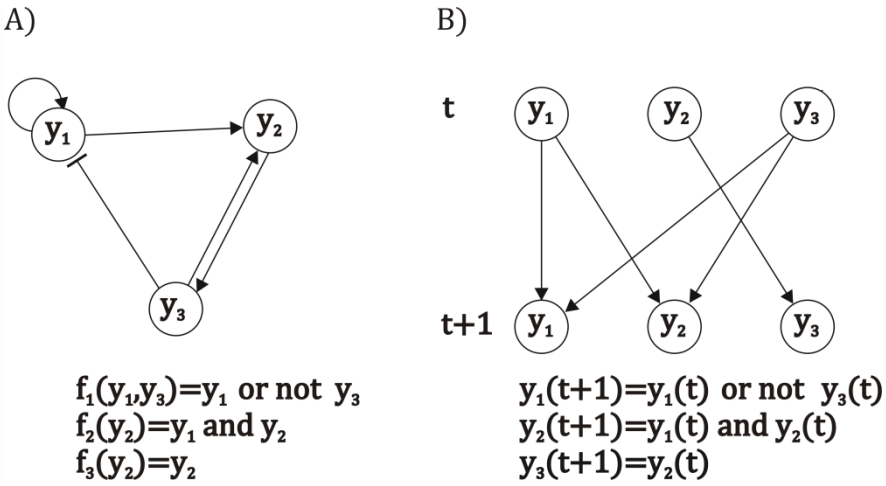


Figure 1. A) An example Boolean network described by Boolean functions. B) Wiring diagram of the Boolean network and corresponding equations for transition between network states t and t+1.

The Boolean variables can only have two logical states 1 (true) or 0 (false) which are connected by logical operators (and, or, not) creating thus a Boolean function. In the Boolean networks each node y_i (gene) is characterized by variables corresponding to levels of gene expression. If it exceeds a certain threshold, the gene is considered "on" otherwise it is "off". Regulatory interaction between genes can be captured in Boolean networks by computing the gene state from the aggregated effect of all its parents (regulators). The state $S(t)$ of the whole network is determined by the states of the nodes (genes) $y_i(t)$ at given time t. To include dynamic to the Boolean models the state of the nodes is synchronously updated at time $t + 1$:

$$y_i(t+1) = f_i(y_{i1}(t), y_{i2}(t), \dots, y_{ik}(t)), \qquad (1.1)$$

where f_i is a Boolean function describing the relationships between the nodes. The state transition between measurements of gene expression at two successive time points corresponds to network transition from the state $S(t)$ to $S(t+1)$.

Boolean networks realistically include dynamics of the biological systems and also can be easily biologically interpreted. However, there are two main limitations of the models. First, the main limitation lies in the fact that real gene expression is not a Boolean variable as it does not follow the rule of only "on" or "off" state but is represented by continuous values. Second, the biological gene networks are typically asynchronous while the model assumes synchrony during the network state transition.

To obtain more realistic networks, basic Boolean models have been extended by distinct approaches such as probabilistic Boolean networks (Shmulevich et al. 2002) or including Post classes functions (Shmulevich et al. 2003).

1.1.2 Bayesian networks

Bayesian networks combine graph and probability theory for a set of random variable $y_i \in Y,\ 1 < i \leq n$ (nodes) and their probabilistic relationships represented by edges (E). The structure of Bayesian network is a directed acyclic graph $G(Y, E)$. If the nodes represent genes, then characteristic variable describing the nodes is the mRNA concentration levels (steady-state genes' expressions) and the edges indicate the influence of regulator y_i (referred as parent) on the expression of gene y_j (referred as child of y_i).

Under the assumption that each variable is independent of its non-descendents, the Bayesian network represents a conditional probability distribution $P(x_i|P_a^G(y_i))$ over a set of random variables y_i where $P_a^G(y_i)$ is the set of parents for each node or, in other words, the set of regulators for each *i-th* gene in the graph G (Figure 2). The feature of the Bayesian networks is that each joined probability distribution can be factorized into a product of conditional distribution:

$$P(y_1, y_2, \dots, y_n) = \prod_{i=1}^{n} P\left(y_i \middle| P_a^G(y_i)\right), \qquad (1.2)$$

In case the node y_j does not have parents the probability distribution is unconditional.

The identification of GRN is a problem of finding the candidate Bayesian network that best fits the gene expression data. To train the Bayesian network from data, a scoring function (model) for the graph evaluation must be selected. An examples of a scoring functions that avoid data overfitting are the Bayesian Information Criteria or Bayesian Dirichlet equivalence (de Campos 2006).

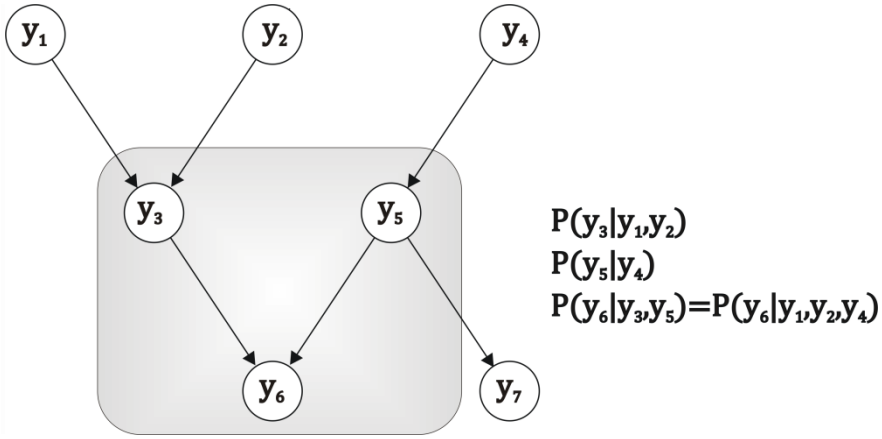


Figure 2. An example of a Bayesian network and conditional probability distributions for gene expression levels y.

The process of parameter fitting consists in finding the best candidates of the conditional probability distributions, usually employing approximations such as Maximum Likelihood Estimation approaches (when all nodes are known), Expectation Maximization algorithms (when some nodes are hidden) or Monte Carlo sampling approach (Lawrence et al. 2009).

The definition of Bayesian networks as directed acyclic graphs imply the major limitations of Bayesian models. They are qualitative and cannot describe feedback loops (cycles); an important feature of real biological GRN (Lee et al. 2002). Hence, the basic form of the model was extended to overcome these limitations by forming Dynamic Bayesian networks where both cyclic networks and the time-series data are included (Yu et al. 2004). In dynamic models a transition from time t to time $t + 1$ is realised by duplication of each node from state at time t to the series of nodes in layers $+1$. Dynamic Bayesian networks thus introduce additional edges between layers.

The main advantage of both Bayesian networks and Dynamic Bayesian networks is that they can be used in cases of missing values or missing variables (Beal et al. 2005). Moreover, Bayesian approach enables incorporating prior information about the system which leads to reduction of number of potential high-scoring networks, a common problem in finding the networks. Together with stochastic simulation approach, they are the only methods that are able to deal with stochastic nature of the cellular processes. The disadvantage of this approach is that due to the limited number of data they can be used only for inference of relatively small networks typically consisting of units, maximum tens of genes. They have been used mostly for inference of partially known networks. They did not prove to be suitable for exploratory analysis, where the knowledge about the network structures is minimal or none.

1.1.3 Ordinary differential equations

Algorithms based on ordinary differential equations describe gene regulations in a general form:

$$\frac{dy_i}{dt} = f_i(y_i, \mu, \theta_i), \qquad (1.3)$$

where $y_i,\ 1 < i \leq n$ stand for transcript concentration of the *i-th* gene, $\frac{dy_i}{dt}$ is the rate of transcript concentration change for *i-th* gene, and $f(.)$ represents the nonlinear function that appropriately describe the relationship between all variables through the set of its parameters θ_i and can additionally include an external perturbation to the system μ .

Ordinary differential equation networks (Figure 3) are deterministic and thus represent an example of quantitative modeling formalism for complex systems. In principle, the models characterize transcript expression rate as a function of other genes' transcript concentrations as described by equation *(1.3)*. The form of $f(.)$ depends on biochemical nature of the process and reflects for example saturation of the gene products which happen at some point in time. Therefore, the functions are generally non-linear and frequently approximated by a sigmoid, Michaelis-Menten or Hill type kinetics (Polynikis et al. 2009). In addition, parameters of such model

kinetics with biochemical background have physical meaning and can characterize real features of the control process in terms of protein-binding rates and affinities and transcription and/or translation rates. The basic model is often extended with additional terms describing effect of other biochemical interaction like mRNA degradation, protein degradation or external perturbations (Gardner et al. 2003). To infer gene network from the measured data, the task is to fit (or estimate) the parameters in the function $f(.)$. When the analytical solution is unknown, parameters can be approximated by numerical methods, however, solving algorithms are usually computationally expensive. A key advantage of models based on ordinary differential equation is their realistic nature; once the $f(.)$ and its parameters are known, the models are able to make predictions about behavior of the network in different state. Disadvantage is that estimation of the parameter matrix for larger networks often fails due to limited number of available data. This limitation can be overcome by piecevise modeling of individual regulator-target gene interactions which, without additional knowledge, can get out of computational abilities due to a large number of possible combinations of regulator-target gene interactions.

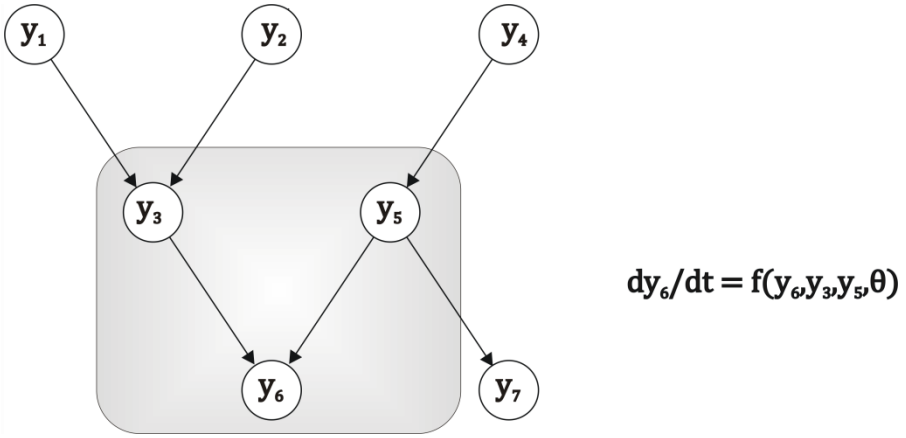


Figure 3. An example ordinary differential equation network. The rate of gene y_6 expression is deterministically modulated by function (f) given by the levels of y_6's direct causal regulators.

1.1.4 Non-parametric approaches to nonlinear dynamical systems

These approaches combine the advantages of Bayesian networks statistics method with nonlinear dynamical systems (Aijo and Lahdesmaki 2009). Generally, the nonlinear dynamic system evolves in time as:

$$\frac{dy_i}{dt} = f(y_i), \qquad (1.4)$$

for $1 < i \leq n$ where $f(.)$ represents an unknown, nonlinear function. Instead of using explicitly parameterized model for $f(.)$ the goal is to obtain character of the function directly from the data by using a non-parametric approach and letting the data "speak" more clearly for themselves. For the case of known regulator (parent) of a particular gene, the task is to identify the relationship between series of outputs y_i from series of inputs $P_a^G(y_i)$. With this assumptions $f(.)$ can be a Gaussian process (Rasmussen and Williams 2006). A Gaussian process, an analogue to the Gaussian distribution, is a probability distribution over all possible functions f that reach the value $f(.)$ at time t and is fully characterized by a mean function μ and covariance function k. The reliability of the model depends on the choice of the covariance function k and the set of its parameters θ which represents the hyperparameters of the Gaussian process. The methods for maximizing the marginal likelihood are used to estimate the set of θ (Rasmussen and Williams 2006). The advantage of this approach is the possibility to estimate real concentrations of the protein regulators instead of predominantly used mRNA concentrations, approaching thus much more the reality of gene expression then the other models do. Disadvantage is that for more than one regulator the model gets excessively complicated and difficult to solve for even medium scale networks.

1.1.5 Information theory based models

The concept of the information theory for biological models is based on adjustment of physical models for information transmission through the noisy communication channel, reviewed in (Tkacik and Walczak 2011). In the genetic

communication channels input regulatory signal $y_i, 1 < i \leq n$ (e.g. regulator expression level) is transmitted through the network of biological processes (e.g. transcription, translation and signaling pathways) to the output effector y_j (e.g. target gene expression level). Generally, the input signal and output effector are nonlinearly related through some noisy relation. The information, conveyed by input signal about output effector level, is transmitted through the system with some level of uncertainty. This uncertainty can be described by entropy:

$$S_i = -\sum_{k=1}^{n} p(y_k)\log\left(p(y_k)\right), \qquad (1.5)$$

where $p(y_k)$ is the probability over a set of variables y_k. Under the assumption of dependency between input signal and output effector, the probability in equation *(1.5)* is conditional and entropy then refers to the number of accessible states of the system but under the constrain that states are unequally probable. This quantity can be derived as mutual information:

$$I_{i,j} = S_i + S_j + S_{ij}, \qquad (1.6)$$

where S_{ij} represents a joint entropy $S_{ij} = S_i + S_j$.

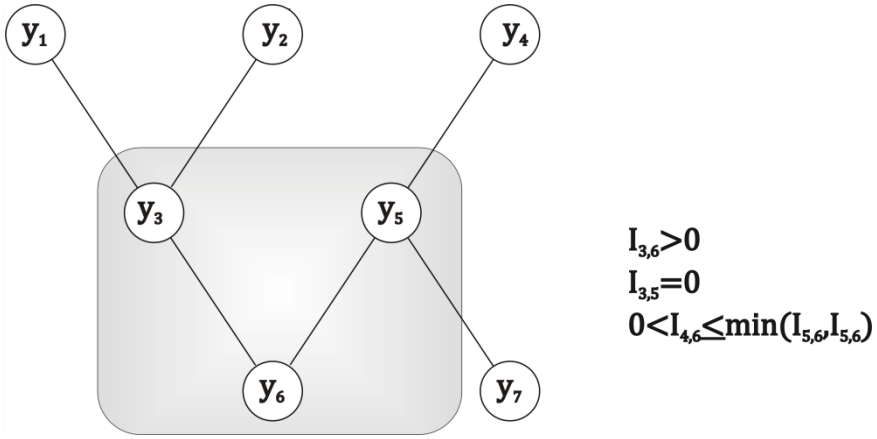


Figure 4. En example information theory based network and corresponding mutual information I for dependent and independent variables.

The mutual information can measure how much on average the uncertainty in output effector level has been decreased by knowing the value of a related input signal. In other words, how strongly the input signals and output effectors depend on each other. Equation *(1.6)* relates the concept of information theory with the measure of interdependency (Figure 4). The $I = 0$ if the variables y_i and y_j (e.g. genes expression levels) are statistically independent. The higher the I, the higher the probability of statistical dependence (e.g. regulation) among genes y_i and y_j.

Regulatory relationships derived by information theory based models thus represent statistical association between genes and do not characterize direct causal interactions. Hence, these relationships must be interpreted in biological sense.

1.1.6 Stochastic equations

The stochastic approach reflects in more details real concentrations of reacting molecules in the cell processes. In the cell only a few molecules are necessary to initiate the process. While assuming the low concentration levels of interacting molecules, the stochastic nature of such systems must be considered, because the influence of fluctuations (noise) in reaction rates is rapidly increasing with decreasing number of interacting molecules. Therefore, unlike in ordinary differential equation based methods where molecular concentrations are assumed to be continuous and evolve in time deterministically, stochastic approaches take into account discrete numbers of molecules and the changes in concentration levels are probabilistic rather than deterministic. The evolution of the system in time can be characterized by chemical master equation (van Kampen 2007) :

$$\frac{dP_k}{dt} = \sum_{l \neq k} (A_{kl} P_i - A_{lk} P_k), \qquad (1.7)$$

where P_k is the probability of the system being in the state k at time t and A_{kl} describes a transition rate from the state k to state l. Equation *(1.7)* characterizes how the probability of being in a certain state changes in time.

Although the stochastic approach describes events more realistically, the computational demands for solving regulatory systems are huge. The way how to overcome computational difficulties involves employing approximations or

stochastic simulation algorithms (Gillespie 1977; van Kampen 2007). Moreover, this approach is limited by requirements for detailed information about reaction mechanisms which are mostly not available. Thus, the pros and cons of employing stochastic models must be considered regarding the desired level of granularity of the regulatory process. It was discussed (Gillespie 2000) that for the large time scales the stochastic effects can be diminished and deterministic models can be used instead.

1.1.7 Neural networks

Originally, the term neural networks described circuits of human nervous cells, neurons. However, nowadays meaning in computational biology refers to a mathematical model of artificial neural networks. The structure of the artificial neural networks consists of neurons (nodes) performing global behavior predetermined by synaptic links (connections) whose synaptic strength is characterized by a weight. In the terms of biological neural systems all networks belong to the class called recurrent neural networks. In the fully recurrent network, all nodes are connected. Hence, in principle, a state of a node is influenced by the activity of all nodes in the network and depends on the weights of individual connections. Subsequently, time-discrete dynamics of the state of *i-th* node $y_i(t + \Delta t)$ at time $t + \Delta t$ depends on the state of the node at time t and the sum of the weighted effects of all n nodes from the network (Figure 5) connected to the given node. Repeating recursively this scheme over whole observed period and with an infinitesimal change of t the rate $\frac{dy_i}{dt}$ of change of the state of a node *i* can be defined as (Bose and Liang 1996):

$$\tau_i \frac{dy_i}{dt} = -y_i + f_i\left(\sum_j w_{ij} y_j - \theta_i\right), \qquad (1.8)$$

where $y_i, 1 < i \leq n$ is a signal strength in the *i-th* node, $\frac{dy_i}{dt}$ represents a rate of the change of the signal in the *i-th* node, functionally dependent on the states of other nodes *j=1,...,n* connected to the *i-th* node. $f_i(.)$ describes a nonlinear transfer

function of the process, w_{ij} represents elements of the weight matrix of all connections strengths, θ_i is an external input to the node i and τ_i is a scaling factor.

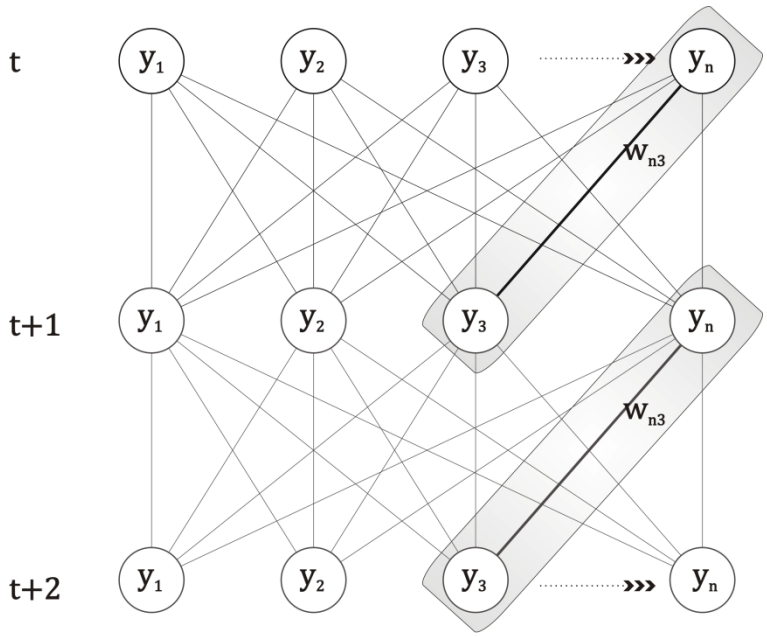


Figure 5. An example neural network. Each node is connected with all other nodes by weighted synapse. The example equation characterizes control of y_3 by y_n, while all other controls are insignificant ($w_{nj}\approx 0$).

Vohradsky (Vohradsky 2001) derived the model for gene expression leading to the same form as equation *(1.8)*. Following the same rationale as above, the state of gene expression (amount of transcribed mRNA) of *i-th* gene at time $t + \Delta t$ is given by the concentration of genes controlling *i-th* gene at time t, reduced by its own decay. For an infinitesimal time increment, formally written

$$\frac{dy_i}{dt} = q_i - d_i, \qquad (1.9)$$

where q_i represents the regulatory effect of all n genes (the nodes of the neural networks) on *i-th* gene expression and $d_i = d_i(y_i)$ is the rate of degradation. The rate of degradation can be described by the first order chemical reaction:

$$d_i = k_{2i} y_i, \qquad (1.10)$$

where k_{2i} is the degradation rate constant. Following equation *(1.8)* the total regulatory effect q_i on *i-th* gene is modulated by a non-linear relation:

$$q_i = k_{1i} f(\sum_j (w_{ij} R_j - b_i)), \qquad (1.11)$$

where the sum represents the effect of regulator R_j on *i-th* gene expression level, given by the strength of connection w_{ij} between *i-th* gene and the regulator and a parameter b_i representing reaction delay. Coefficient k_{1i} characterizes the maximal rate of expression and non-linearity of the interactions describes function $f(.)$

To interpret the transfer function $f(.)$, the nature of the transcription event must be considered. The transcription initiation in the cell depends on the probability of regulator binding to the promoter sequence of the gene. Then, the rate of transcription is given by the strength of regulatory effect of the regulator to particular gene ($\sim w_{ij}$). The probability of regulator binding depends on the number of regulator molecules; in case of low regulator concentration the regulatory effect and thus gene expression rate are small in general (Figure 6). Increasing amount of regulator molecules evokes enhancement also in the target gene expression rate. For certain regulators' concentration, the gene expression rate stays almost constant and does not increase with increasing number of regulatory molecules any more due to the saturation of the gene promoter. The non-linear relationship between regulator concentration and gene expression rate can be described by sigmoid function (Veitia 2003):

$$f(z) = \frac{1}{1 + \exp(-z)} \qquad (1.12)$$

Combining equations *(1.9)* to *(1.12)* resulting model for gene regulatory interaction has a form:

$$\frac{dy_i}{dt} = \frac{k_{1i}}{1 + \exp\left[-\left(\sum_j w_{ij} R_j + b_i\right)\right]} - k_{2i} y_i, \qquad (1.13)$$

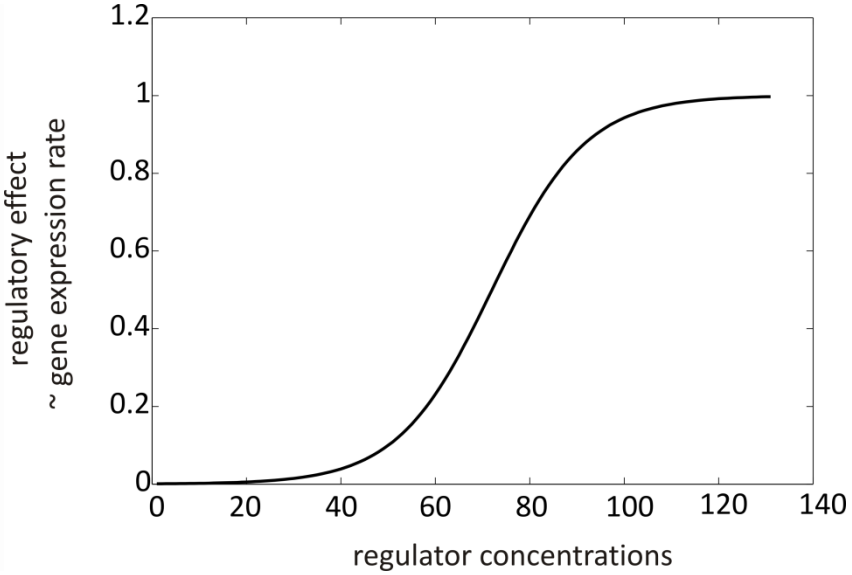


Figure 6. An example of the regulatory course. The regulatory effect on the gene expression rate depends on the amount of regulator molecules. The relationship is non-linear and has a sigmoid character (Veitia 2003).

The same rationale in the derivation of the equations *(1.13)* and *(1.8)* and their comparison shows identical principle for both the recurrent neural network and a transcriptional process model. The only difference is that in equation *(1.8)* temporal change of the state of the *i-th* node is scaled together, while in equation *(1.13)* it is scaled individually for each term according to the biochemical principles of transcription. In the model *(1.13)*, individual steps of the transcriptional process were replaced by the observed behavior of promoters initiation and saturation. From the chemical kinetics point of view, all individual factors involved in the transcriptional process are available in excess and therefore do not influence the rate of transcription. The only thing which remains undefined is the shape of the function $f(.)$. Here it was assigned a sigmoid shape; the shape that was observed experimentally and it is also supported by theoretical considerations (Veitia 2003).

Mathematically, the model based on recurrent neural networks *(1.13)* lead to the set of ordinary differential equations, similar as mention above (paragraph 1.1.3). The advantage of the neural networks approach consists in derivation process, which reflects more realistically the nature of transcription and gives the parameters biologically interpretable meaning.

1.2 Experimental methods

Extension of high-throughput technologies in recent years enables performing experiments at a large scale rather than at a single-gene level. Diverse experimental approaches in systems biology used for comprehensive analysis of biological systems are commonly referred as “omics”. The shared feature for all omics platforms is the emphasis to measure the whole response of the system on the global scale in terms of particular treatment or physiology. The most progressives are the fields of genomics, transcriptomics, proteomics and metabolomics but there are more other subdisciplines. As the names suggest, genomics analyzes DNA structure and function, transcriptomics studies temporal RNA pool, proteomics analyzes proteins structure and functions on large scale and metabolomics monitors cell physiological activity by measuring changes in metabolic pathways.

From those mentioned, two omics methods, transcriptomics and quantitative gel-based proteomics, were used here for the data gathering. In transcriptomics approach, DNA microarrays are employed to qualitatively and quantitatively determine amounts of gene expression products - mRNA molecules. The principle of DNA microarray method is described in paragraph 1.2.1.

Quantitative proteomics of complex protein sample ideally identify and quantify all individual proteins from a mixture. To identify a particular protein, various mass spectrometry techniques are used. However, quantification of individual proteins from complex mixture is less straightforward. The main reason staying behind the difficulty of protein measurements is, that unlike RNAs, proteins have enormous variability in forms, composition and physical-chemical properties. Therefore it is challenging to capture all possible features by a single technology with reasonable accuracy and reproducibility. Nevertheless, several strategies for inferring protein quantity are used such as two-dimensional (2D) gel electrophoresis, chromatography and gel-free methods based on special sample labeling and further evaluating by high-resolution mass spectrometry techniques. Even though advances in the technology and bioinformatics nowadays enable to analyze hundreds or even thousands of proteins, definitive quantification of the entire proteome of a complex protein sample is still impossible. In the paragraph 1.2.2 the fundamentals behind the employed 2D electrophoresis based proteomics are described.

1.2.1 DNA microarrays technique

DNA microarrays represent a high-throughput method that allows detection and quantification of thousands of genes simultaneously. The microarray chip is a solid surface, usually glass, with tens of thousands of DNA gene oligonucleotide sequences chemically attached at specific positions. The immobilized gene sequences are called probes and the cluster of probes, originated from the specific gene, creates a spot. One microarray usually contains tens of thousands of spots enabling to quantify thousands of expressed genes simultaneously.

The principle of the microarray technique lies in the hybridization process when labeled sample complementary DNA (cDNAs) specifically bind to the complementary probes on the chip.

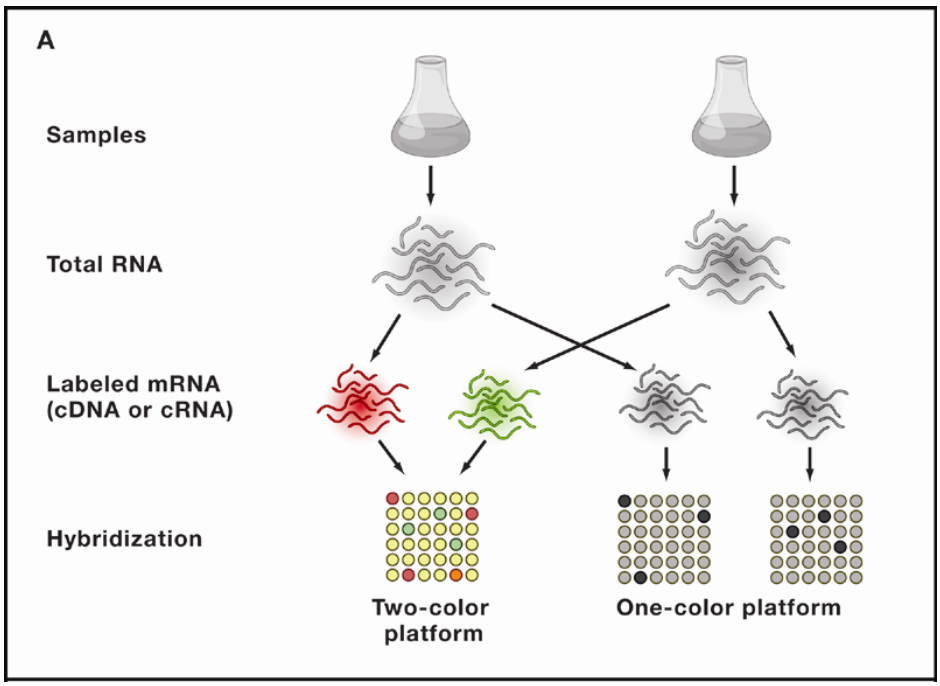


Figure 7. The scheme of microarray experiment for two-color platform and one-color platform. Reprinted from (van Bakel and Holstege 2008).

Typical experimental design compares gene expressions of two different samples (e.g. wild type and mutant or a sample and a common reference in time series experiments). Thus, the obtained quantity of transcripts always characterizes relative changes in mRNA amounts between two samples rather than absolute

concentration levels. The expression of specific gene in the cell means that the gene is actively transcribed into mRNA; the unique complementary molecules to the gene nucleotide sequence. The first step of the procedure involves isolating whole transcriptome (mRNA) from the cells (Figure 7).

Next, the total mRNA is reversibly transcribed into cDNA and labeled. The labeling depends on the type of experimental setting; in the two-color platform each sample is labeled with a different fluorescent dye such as Cy3 (green Cyanine dye) and Cy5 (red Cyanine dye) and subsequently hybridized to the same microarray chip. In the case of one-color platform, only one fluorescent dye is used but the samples are hybridized to separate microarray chips (Figure 7). The hybridized probes with cDNA samples on the microarray are then scanned to obtain fluorescence intensities and locations of the labeled spots. The intensity of the spot is proportional to the amount of bound sample cDNA molecules, representing amount of mRNA transcript. To correct for sequence- and region-dependent biases in the experiment, intensities from two samples are compared to obtain relative gene expression levels.

The main advantage of the method is in determining expression levels simultaneously for thousands of genes. Moreover, for organisms with relatively small genome, like bacteria, the probes covering whole genome can be spotted on a single microarray chip.

1.2.2 Two-dimensional electrophoresis

The 2D electrophoresis is a gel-based method suitable for separation and quantification of protein mixtures. The method takes advantage of combining two different separation procedures to increase a resolution of proteins of similar physical-chemical properties (like size, mass and charge) and thus enhances the probability of separation because it is less likely that two proteins share both distinct properties.

In the first step, proteins are sorted according to their intrinsic charge by isoelectric focusing (Figure 8A). The procedure proceeds in a gel strip with a stable pH gradient under applied electric field. Each protein charge interacts with external electric potential resulting in migration in the gel. The protein intrinsic charge depends on amino acid content and surrounding pH; every protein is characterized by isoelectric point, which is the pH value where it has zero net charge and therefore

does not react to the external electric field and stays immobile in the strip. Eventually, in the first dimension procedure, each protein remains fixed in the position in the gel strip where the pH corresponds to its isoelectric point.

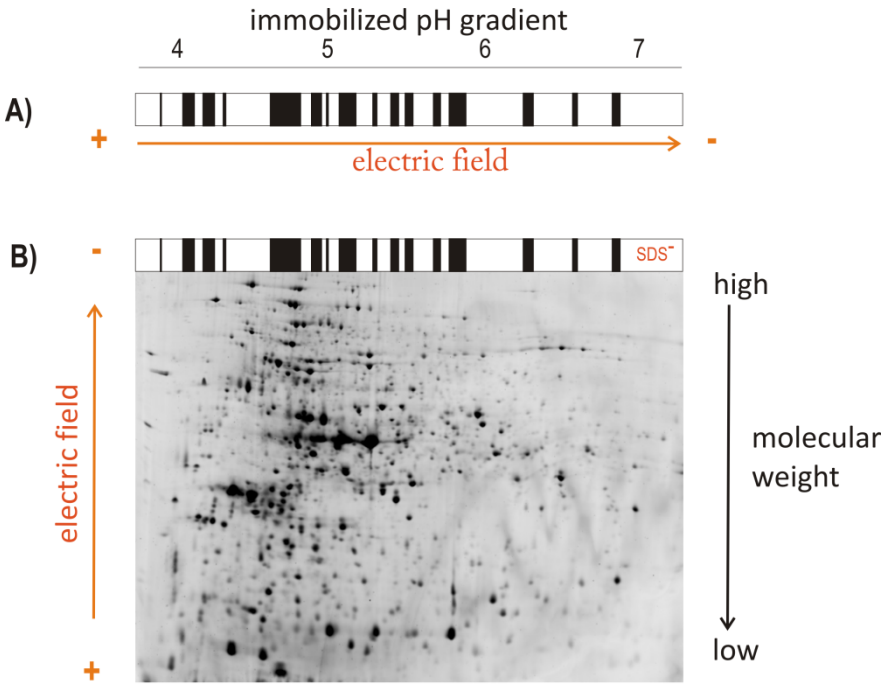


Figure 8. A scheme of two-dimensional gel electrophoresis. A) Isoelectric focusing separates proteins by charge on immobilized pH gradient gel strip (first-dimension separation). Proteins migrate in external electric field to their isoelectric point. B) Gel electrophoresis separates pre-sorted proteins according to their molecular weight (second-dimension separation). Denatured proteins are coated with negatively charged SDS molecules. Proteins' charge is proportional to their molecular weight and thus proteins terminate migration in the external field according to their mass. The example gel is from *S. coelicolor* germination made in our lab. Proteins were stained by SYPRO Ruby.

In the second step, proteins are separated by their mass. Primarily, proteins fixed in the gel strip are denatured and coated with sodium dodecyl sulphate (SDS); the negatively charged compound. The number of SDS molecules bound to each protein is proportional to the protein mass. Hence, the protein's charge in the second dimension of the electrophoresis corresponds to its mass. Next, the second electrophoresis is run (Figure 8B). The gel strip is placed to the edge of the

rectangular thin gel and electric field is applied. The arrangement allows the SDS-charged proteins to migrate through the gel (2D gel) in the direction that is right angled to the original isoelectric focusing moving direction. The length of the protein trajectory depends on the protein charge and mass, respectively. The smaller the protein, the longer the migration in the gel. Each protein creates a circle shaped spot in the gel. To visualize proteins in the gel various staining techniques are used (such as silver staining, Coomassie blue staining, fluorescent dyes staining or prior protein radioactive labeling).

Stained 2D gel is then scanned to determine intensities of individual protein spots. The intensity of the protein spot is proportional to its quantity. For protein spot quantification and 2D gel images comparison a special software is used.

In addition, the protein spots of interest can be extracted from the 2D gel and identified by mass spectrometry.

Although it is still not possible to quantify all proteins from the mixture, the 2D gel electrophoresis represents a powerful method for protein separation and quantification.

1.3 Streptomyces

Streptomyces species belong to the large group of *Actinomycetes*. *Streptomyces* are gram-positive, soil dwelling bacteria with a G-C rich genome. They have one of the largest genomes of the bacterial kingdom (8-10Mbp) and their chromosome, unlike in most bacterial species, is linear. The large chromosome implies enormous capacity for regulation, enabling cell differentiation and adaptation in a range of environmental conditions. Due to the complex life cycle and ability to produce antibiotics, *Streptomyces* serve as model organisms to study diverse aspects of the cell biology. During the complex developmental cycle, *Streptomyces* differentiate into several types of cells (Figure 9). At the beginning, the single ovoid spore germinates and forms vegetative mycelium by filamentous growth and branching. In the next phase, aerial hyphae rise to create spore chains which eventually evolve into individual exospores. Thus, the single bacterial spore can generate colonies of new individuals. Different cell types have not only different morphology but also physiology and metabolism, govern by larger number of regulatory genes. During the differentiation, number of secondary metabolites and antibiotics is synthesized.

Many of these compounds produced by *Streptomyces* species have been used in medicine and agriculture. Hence they have become valuable organism for industrial purposes. Widely studied model organism, representing the genus of *Streptomyces,* is *Streptomyces coelicolor*, whose genome has been sequenced and annotated. Moreover, the whole-genome microarray chips are available for *S. coelicolor*.

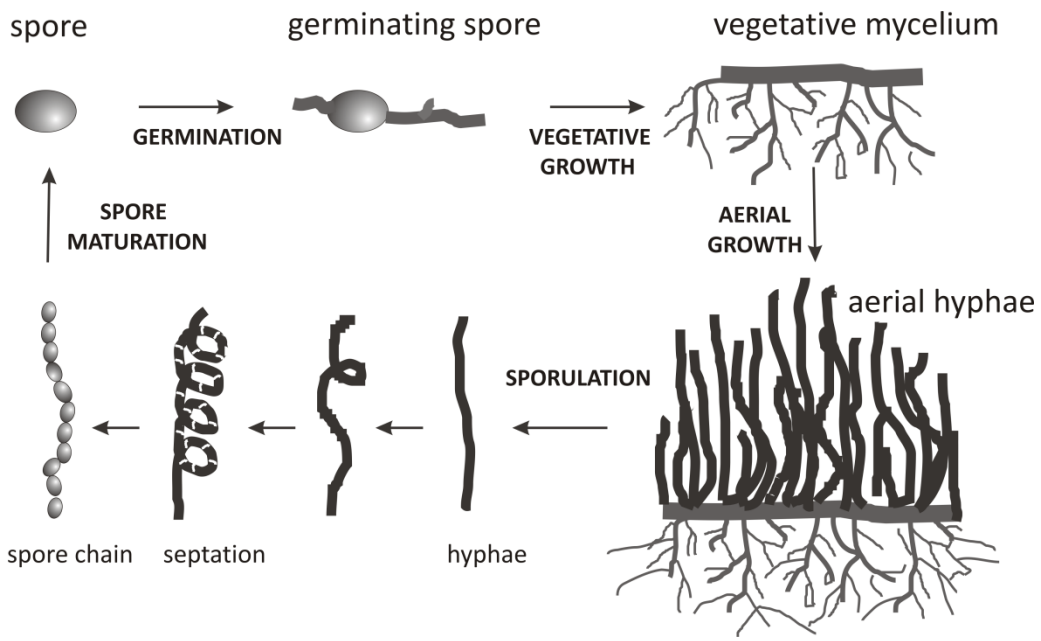


Figure 9. The life cycle of *Streptomyces coelicolor*. Under favorable conditions germination is triggered in a single spore followed by vegetative growth. Next aerial hyphae rise. During the sporulation phase, aerial hyphae are partitioned by septa to produce chains of unigenomic spores, which can be dispersed and start a new cycle.

1.3.1 Germination

S. coelicolor spore is an ovoid cell with a thick inner peptidoglycan wall and thin outer layer. The outer layer is composed of rodlin and chaplin proteins organized into basketwork-like structure on the spore surface (Claessen et al. 2004). The spore wall composition and structure provide protective coat enabling to resist physical or chemical treatments and protect cell content against unfavorable external conditions. Inner spore environment is characterized by low water content and increased presence of immobile protein aggregates (Cowan et al. 2003; Mikulik et al. 2002). It was suggested, that dormant spores contain many compounds synthesized during the

sporulation phase such as hydrolases (Haiser et al. 2009) for a cell wall lysis to allow influx of external nutrients, chaperones (Bobek et al. 2004) for assistance in protein assembly and energy sources e.g. trehalose (Ranade and Vining 1993), polyphosphates (Ghorbel et al. 2006). It was reported, that spore also possesses stable mRNA to provide templates for early translation in germination (Mikulik, et al. 2002). Typical for dormancy is only minimal metabolic activity.

The transition from dormancy to active metabolism is called germination. Under favorable external conditions (water milieu, nutrients) germination is initiated. Immediately, after germination initiation, transcription and translation machinery begin to synthesize *de novo* molecules (Bobek, et al. 2004). Bobek et. al. studied association of chaperones with ribosomes and proteins modified by reversible phosphorylation during germination. The role of chaperones in ribosome activation and protection of nascent proteins from aggregation in early germination was suggested.

Germination rate was shown to be influenced by signaling molecule cyclic AMP (cAMP) (Susstrunk et al. 1998). The studies (Derouaux et al. 2004a; Derouaux et al. 2004b; Susstrunk, et al. 1998) with mutations in cAMP-binding proteins *in S. coelicolor* revealed the reduction in the germination rate. The central regulatory role in all developmental steps was suggested for Crp (cAMP receptor protein) (Piette et al. 2005). In addition, mutation in *crp* led to decrease of the level of important cell wall hydrolase (SCO5466), which is considered to be an essential compound responsible for peptidoglycan wall degradation in the germinating cell. Generally, crucial role of hydrolytic enzymes, such as peptidoglycan hydrolases RpfA (SCO3097) and SwlA (SCO1240), in *S. coelicolor* growth and development was proposed (Haiser, et al. 2009). After metabolism recovery, the first DNA replication occurs and overall transcription and proteosynthesis are enhanced (Mikulik, et al. 2002).

Physiologically, spores during germination swell and get bigger. In the position marked with SsgA protein during sporulation phase, germ tube emerges (Noens et al. 2005). The germ tubes rise from the inner wall of spore through the outer wall and are microscopically visible. Crucial compound for further hyphae growth is coiled coil protein DivIVA (SCO2077), which determines polar tip growth by association with cell wall biosynthetic apparatus (Flardh 2003a; Flardh 2003b). Subsequent

development to vegetative branching mycelia continues with further protraction of the tubes.

Although several observations have been made on germinating spores, the molecular and regulatory mechanism underlying the triggering of the process remain still unknown.

2 Objectives of present study

Contemporary views perceive living cells as a complex interacting system. The holistic concept of biological systems evolved in the wide field of systems biology, integrating experimental and computational approaches. The main event that has influenced progression in the field is the expansion of high-throughput techniques generating large amounts of quantitative data about particular systems. Subsequently, demands on development of computational methods for analysis of such high-throughput datasets to infer, identify and model relations between the molecules and pathways in the cells have risen. The vision of the systems biology community is to create a "virtual cell" via computer simulations that would realistically model the interactions between thousands of cellular components to provide comprehensive information about fundamental processes in the cell, with future potential applications in diseases research and pharmaceutical industry.

One of the major focuses in the systems biology approach nowadays, is identification of gene regulatory networks through a combination of both experimental biology and computational methods. Suitable model system for identification of GRN should be reproducible, synchronous and ideally with a defined origin. An example of such system is the process of *Streptomyces* germination. During germination, spores undergo a developmental transition from dormancy to vegetative growth. Because spores contain a consistent pool of protein and RNA molecules, the development always begins from the well-defined initial state. Despite the fact that germination influences further growth and differentiation, it has not been systematically studied and thus the molecular basis or principal metabolic and regulatory mechanisms are not known so far.

The objective of the present project was the identification of genes and metabolic processes associated with initiation and other phases of the *S. coelicolor* germination by employing a systems biology approach. The aim of the experimental part of the project was the monitoring of the quantitative changes in gene products concentrations over time during the progression of germination to obtain time series data. The goal of the computational part of the project was analyzing the experimental time series data by applying numerical models and statistical analysis to reveal processes controlling *S. coelicolor* germination.

Reflecting the above motivation, aims of the project are summarized as follows:

- To generate quantitative time series data from the initial phase of *S. coelicolor* germination by monitoring concentration level profiles of mRNA, accumulated proteins and newly synthesized proteins.

- To reconstruct genetic networks under transcriptional control of sigma factors through applying the kinetic numerical model of gene expression to measured microarray data.

- To identify biochemical processes associated with *S. coelicolor* germination by processing the proteomic time series and identifying selected proteins by mass spectrometry.

- To compare proteome and transcriptome kinetic responses during *S. coelicolor* germination and employ computational methods for analyzing global features of the system.

The concept of the thesis is based on summary of main results (Chapter 3) reported in following publications:

Paper I (Appendix A)
STRAKOVA, E., A. ZIKOVA and J. VOHRADSKY. Inference of Sigma Factor Controlled Networks by Using Numerical Modeling Applied to Microarray Time Series Data of the Germinating Prokaryote. *Manuscript in review*

Paper II (Appendix B)
STRAKOVA, E., J. BOBEK, A. ZIKOVA, P. REHULKA, O. BENADA, H. REHULKOVA, O. KOFRONOVA and J. VOHRADSKY . Systems Insight into the Spore Germination of *Streptomyces* coelicolor. J Proteome Res, Jan 4 2013, 12(1), 525-536.

Paper III (Appendix C)

STRAKOVA, E., J. BOBEK, A. ZIKOVA and J. VOHRADSKY. Global Features of Gene Expression on the Proteome and Transcriptome Levels in S. *coelicolor* during Germination. PLoS ONE, *accepted July 2013*

A more comprehensive description of the results and used experimental methods can be found in the attached manuscripts (Appendices 8). The papers are referred by the Roman numerals in the text. The authors' contributions to the study are specified at the end of Chapter 3.

3 Results

3.1 Experimental design and data collection

To obtain comprehensive insight into *S. coelicolor* germination, changes in transcriptome and proteome were measured during the initial phase of developmental process. Details about spore cultivation and used chemicals are described in Paper II. For germination initiation in liquid media, dormant spores were subjected to mechanical disruption of the outer coat and subsequent 10 min heat shock at 50°C, a procedure which is considered standard in the field (Kieser et al. 2000). Both procedures were reported to activate germination and boost synchrony of the population. Thus, during observed initial 5.5 hours of *S. coelicolor* germination, the development of spore population was relatively synchronous (microscopically inspected) and arose germ tubes became visible. The physiological change through the process is shown in electron microscopy image in Figure 10.

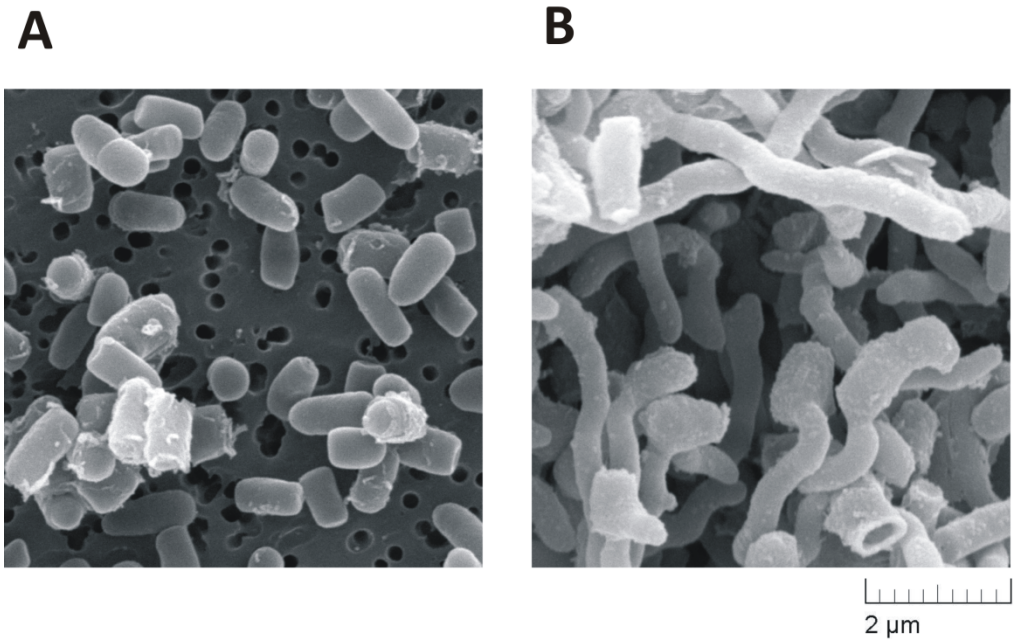


Figure 10. Electron microscopy images of *S. coelicolor* spores at primary magnification of 30 000 times A) Dormant spores (T Dorm). B) Germinating spores, 5.5 hours after germination initiation with grown germ tubes. The electron microscopy images were prepared by Oldrich Benada and Olga Kofronova, Institute of Microbiology, Czech Academy of Sciences.

For the time series analysis and modeling, it is advantageous to monitor as much time points as possible; however, the length of the time intervals is limited by both experimental availability and costs. The samples were collected in 30 min intervals within the process, which altogether gave measurements at 13 time points. To eliminate experimental errors, samples from each time point were examined in replicates originating from different cultivations. The initial time point of time courses was a sample from the dormant spores (T Dorm). The next sample (T0) was obtained immediately after germination initiation (after mechanical disruption and 10 min heat shock, experimental details are described in Paper II). Subsequent samples were obtained every 30 min up to 5.5 hours (referred as T0.5 – T5.5).

3.1.1 Gene expression time series inference

The method for transcriptome (total RNA) isolation was optimized to gain enough high quality RNA samples from spores for microarray measurements. RNA was isolated from germinating spores at 13 time points in three replicates. Each sample was then examined for RNA levels on separate two-color microarray chip (Agilent) by Oxford Gene Technology (Oxford, United Kingdom). As a common reference a mixture of RNA from all examined time points was used. The probes on the microarray chip covered entire *S. coelicolor* genome (7825 genes) as well as probes tilled across the whole genome. Finally, whole experiment provided data from 37 microarray chips (three replicates for each time point except for T1 and T2.5 were only two microarray replicates fulfilled quality control check). Entire dataset was normalized and microarray spot intensities were averaged across replicates to obtain gene expression profiles, representing temporal changes in mRNA expression kinetics. Details for experimental procedures, data normalization and gene expression profiles inference are described in Paper I.

The microarray data discussed in this thesis have been deposited[1)] in NCBI's Gene Expression Omnibus (Edgar et al. 2002) and are accessible through GEO Series accession number GSE4441.

[1)]http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE 44415

3.1.2 Protein expression time series inference

Total soluble protein was isolated from germinating spores at the same time points as described above for RNA samples (13 time points, 3-5 replicates for each time point). Proteins were separated and quantified by 2D gel electrophoresis. We examined not only proteins accumulation, but also proteins synthesis. To detect *de novo* expressed proteins, spores were incubated in the presence of a radioactively labeled amino acid (^{35}S cysteine-methionine) for 30 min. During the *in vivo* radioactive pulse, radioisotope incorporated into newly synthesized proteins. Thus, each gel image provided two types of information. The fist gel image corresponded to total protein accumulation (stained by fluorescent dye SYPRO Ruby – further referred to as "Sypro"). The Sypro intensity of the spot was proportional to the protein concentration in the cell. The second gel image reflected only *de novo* synthesized proteins which were radiolabeled. Here, the spot intensity was proportional to the level of protein translated during the 30 min radiolabeling period. Altogether 104 gel images were compared and analyzed in a single high level matchset in PDQuest software (Bio-Rad) to obtain individual protein's accumulation and synthesis time series. We detected protein profiles for 782 protein spots. From the detected set, 251 individual protein species were identified by mass spectrometry. Details for experimental procedures, gel intensities normalization and protein profiles inference are described in Paper II. The spot intensity values are listed in Supplementary Table 1 to Paper II on the attached CD.

3.2 Transcriptional regulation of germination

In bacteria, crucial regulators of transcription initiation are sigma factors. Sigma factors are proteins that are able to recognize and bind to the target gene promoter region and in association with RNA polymerase start gene transcription into mRNA. The genome of *S. coelicolor* has 65 annotated sigma factors which is one of the largest numbers in bacterial kingdom (Bentley et al. 2002). The number of sigma factors encoded in the bacterial genome is proportional to the developmental complexity of the organism. For comparison, there is only 1 sigma factor in a simple parasitic bacterium *Mycoplasma genitalium*, 7 sigma factors in more evolved

Escherichia coli and 18 sigma factors in endospore-forming bacterium *Bacillus subtilis*. Consequently, due to the generally higher portion of regulatory proteins in *S. coelicolor* (approximately 12% of entire genome (Bentley, et al. 2002)), higher complexity in the regulatory system can be anticipated.

From the perspective of gene regulatory networks, sigma factors represent the essential nodes of the network that control further interactions and processes in the cell. A key task of this study was to identify sigma factors significant for directing germination and recognize their possible target genes. To reveal such regulatory interactions, a kinetic model of transcriptional control based on the recurrent neural networks was used. The computational model (introduced in paragraph 1.1.7) is represented by the general equation *(1.13)* for the fully connected network

$$\frac{dy_i}{dt} = \frac{k_{1i}}{1 + \exp\left[-\left(\sum_j w_{ij} R_j + b_i\right)\right]} - k_{2i} y_i, \qquad (1.13)$$

(where $y_i, 1 < i \leq n,$ is a concentration level of the *i-th* node, $\frac{dy_i}{dt}$ represents a rate of the concentration change, $R_j, 1 < j < m,$ is the concentration of *j-th* regulator, w_{ij} represents the strength of regulatory connection, b_i is a reaction delay, k_{1i} characterizes the maximal rate of expression and k_{2i} is degradation rate constant).

The equation *(1.13)* relates the expression levels of the regulators (the sigma factors in this study) to the expression levels of the target genes. The relation depends on the regulator concentration kinetics profile. To determine transcriptional control from the kinetic perspective, the essential features are kinetic trends (shapes of the temporal expression profiles) of the genes and regulators rather than their absolute expression values. The strength of the control is defined by elements in a weight matrix w.

In principle, reconstruction of the weight matrix w in the model *(1.13)* allows identifying of mutual interactions among genes and inferring gene regulatory networks. However, in the real systems, where only expression kinetics are known, optimizing parameters of the whole network under the assumption of fully connected network, can lead to a set of minimal values w_{ij} in the weight matrix w, giving ambiguous results impossible to interpret (Vu 2005 doctoral thesis). In order to avoid this hurdle, the problem of w reconstruction was particularized to model individual

interaction between one regulator R_j and one target gene y_i. Then, the model *(1.13)* takes a simplified form

$$\frac{dy_i}{dt} = \frac{k_{1i}}{1 + \exp\left[-\left(w_i R_j + b_i\right)\right]} - k_{2i} y_i, \qquad (1.14)$$

The task became a problem of fitting the measured gene expression series y_{m_i} by parameterized function y_i, which form is given by an ordinary differential equation *(1.14)*. The y_i in the equation *(1.14)* can be evaluated numerically; in this work was used function ode45 in Matlab software (MATLAB 2010), based on an explicit Runge-Kuta formula. As a criterion of goodness of a fit between measured y_{m_i} and computed y_i expression profiles the Pearson correlation coefficient c_i was used. The coefficient can gain values in interval [-1,1]. The closer is the coefficient to 1, the better is the linear correlation between the profiles implying that the computed profile y_i has a very similar kinetic trend as the measured expression profile y_{m_i}.

In order to find weight w_i of the interaction together with a set of other parameters k_{1i}, k_{2i}, b_i, of equation *(1.14)*, which would result in a maximal possible correlation coefficient, objective function

$$F_i(k_{1i}, k_{2i}, b_i, w_i) = 1 - c_i(k_{1i}, k_{2i}, b_i, w_i), \qquad (1.15)$$

was minimized. As an optimization procedure the simulated annealing (Spall 2003) was used; a method avoiding trapping in local minima during the iterative optimization.

The goodness of fit (c_i) between experimental and computed expression profiles was evaluated and those interactions that satisfied the criterion were considered as possible (for details see methods in Paper I). Assembling all individual interactions, a possible form of the regulatory network controlled by the sigma factors was reconstructed.

Optimization of each sigma factor-target gene model parameters gives opportunity to simulate kinetic behavior of a target gene as a response to the kinetics of the sigma factor. Having parameters for all interactions would allow simulating kinetic behavior of the whole genetic network. The modeling of kinetic behavior of

the whole genetic network would consequently enable to perform virtual experiments such as gene deletion, overexpression and altered sigma factors kinetics. Simulations can be used to predict behavior of the network and suggest experiments for wet lab verification.

Using measured gene expression time series from *S. coelicolor* germination, we computationally modeled transcriptional regulatory relationships for 45 highly expressed sigma factors. (Identification of highly expressed sigma factors is specified in Paper I). Transcriptome was analyzed on a global scale by substituting individual target genes expression profiles by typical kinetic trends. It means that instead of modeling one-by-one interaction for all combinations of sigma factor – target gene, we rather worked with 42 typical kinetic profiles. Each "core profile" (as these 42 typical trends are called in Paper I) characterized a cluster of genes with a similar expression kinetic. This simplification is possible for two reasons. First, genes controlled by the same sigma factor are expected to have similar expression kinetics. Second, due to the experimental errors and biological variability, each expression profile has a certain confidence interval. Grouping genes to clusters with similar kinetic (expression) profiles within a certain confidence interval allows selecting genes which are potentially controlled by a single regulator. To identify kinetically possible regulations, the transcriptional model *(1.14)* was applied to each pair sigma factor – target cluster core profile. We identified possible interactions between sigma factors and gene clusters for 29 sigma factors (Figure 11 and Figure 12) out of 45 suggested as highly expressed. For the remaining 16 sigma factors no regulatory interaction among them and core cluster profiles was found. Suggested target genes belonging to individual clusters are listed in Supplementary Table 2 to Paper I provided on the attached CD.

The kinetically plausible regulations were proposed for several known sigma factors, as well as for many annotated sigma factors whose function and role have not been previously studied (marked by SCO number Figure 11).

Figure 11 shows all equally probable regulatory alternatives (when more arrows are pointing to one cluster in Figure 11) based on kinetic aspects and cannot be misinterpreted with simultaneous regulations by multiple sigma factors because used model (*1.14*) does not cover simultaneous or competitive regulation of more sigma factors. The Figure 11 was redrawn into more detailed Figure 12 where all proposed target genes from the kinetic clusters are shown.

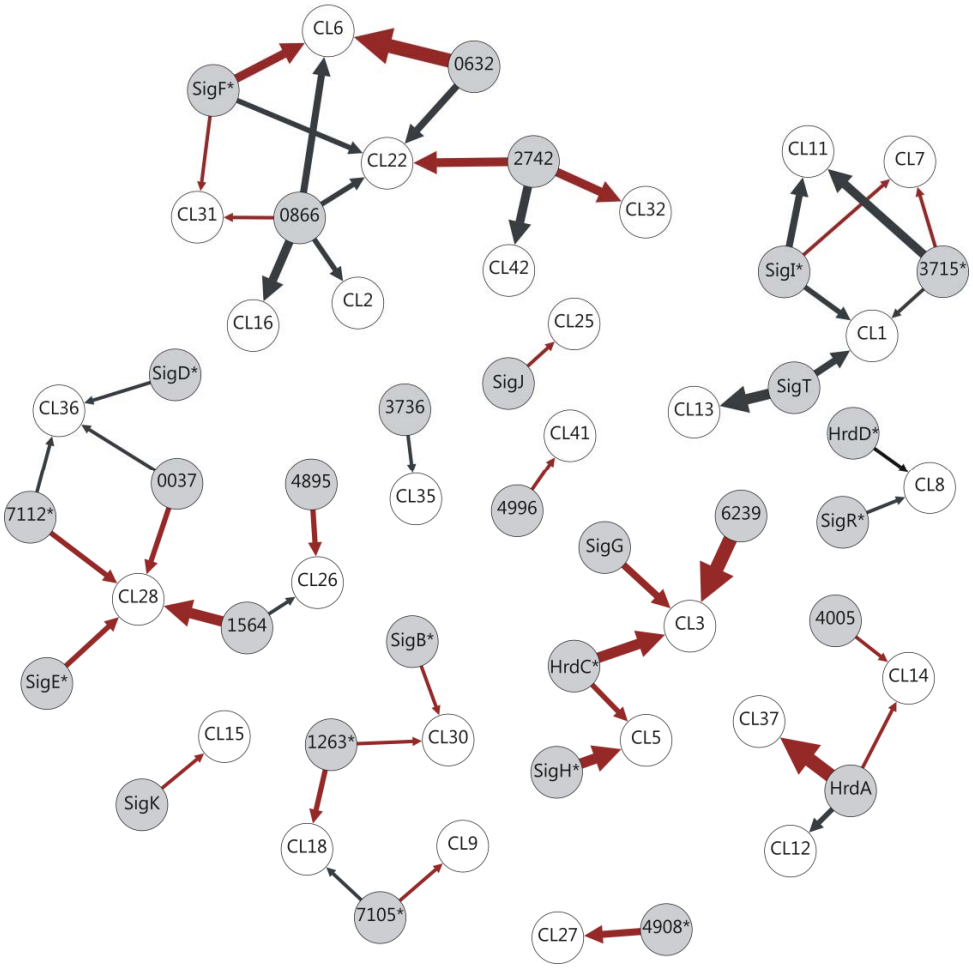


Figure 11. Suggested transcriptional regulations between the sigma factors (grey nodes) and the clusters/groups of genes with similar profiles (white nodes). The black arrows mark trivial regulations, while the red arrows mark nontrivial regulations. The thickness of the arrows is proportional to parameter *w* of the model and corresponds to the strength of the regulatory effect. Sigma factors are marked by name or gene SCO number; clusters are marked with cluster number. Asterisks designate sigma factors, whose overall expression was among 5% of the most expressed genes (details in Paper I).

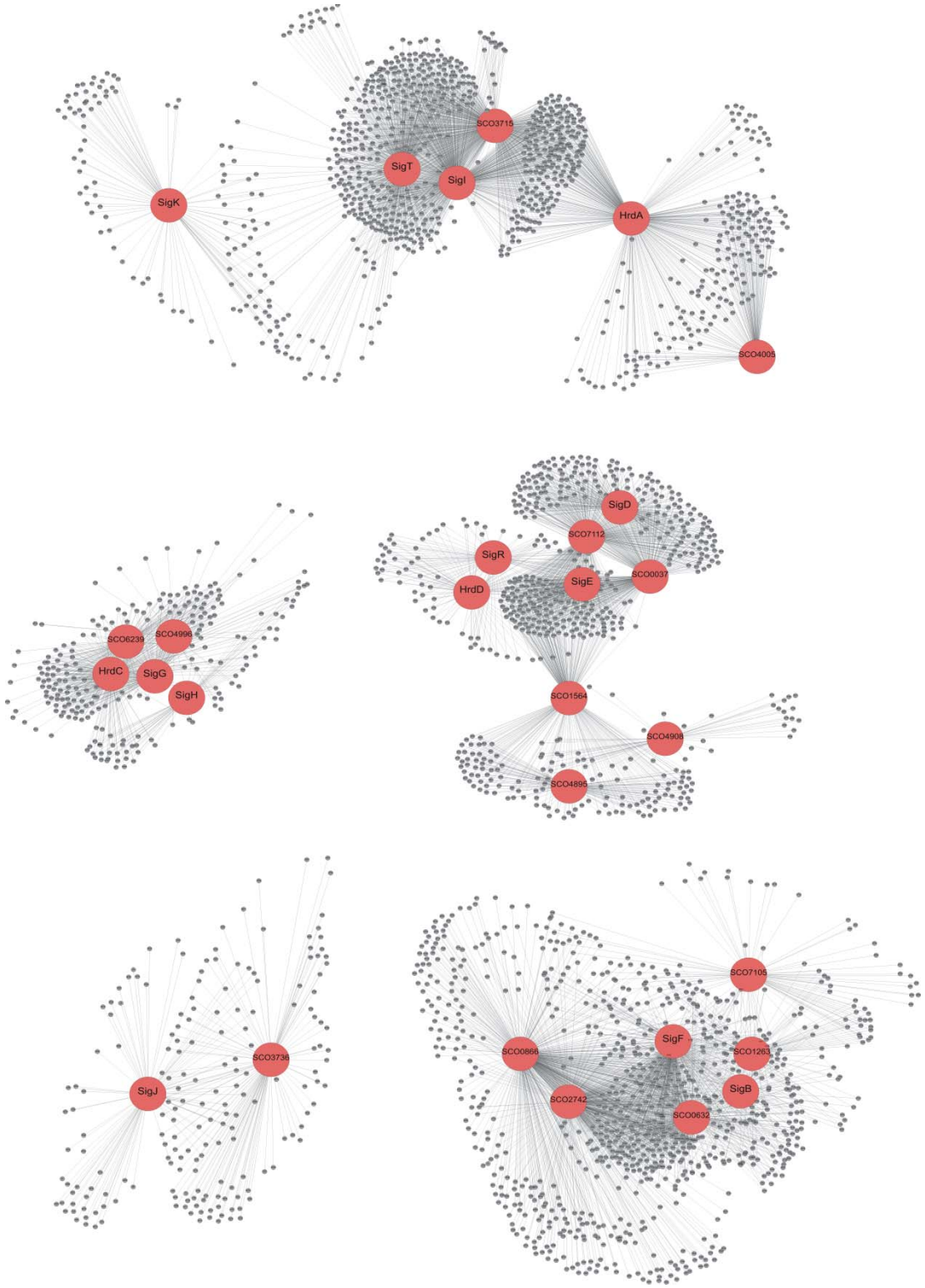


Figure 12. The suggested scheme of the transcriptional regulatory network under the control of sigma factors (the red nodes). The individual target genes from the kinetic clusters are represented by the gray circles. The figure is a more detailed version of Figure 11 and is also provided on the attached CD in EPS format.

Additionally, assigned functions of the gene members in the clusters were considered and significantly overrepresented functional groups were identified (details in Paper I). The idea was to characterize individual kinetic clusters by dominant gene functional groups. As the sigma factors controlling the kinetic clusters were proposed above, we can extrapolate that the metabolic processes characteristic for individual kinetic cluster were also under control of these sigma factors. Thus, several clusters were characterized by enriched functionally related genes (Figure 13). In Figure 13 metabolic processes (white nodes) which we suggested to be under particular sigma factor (gray nodes) control are illustrated.

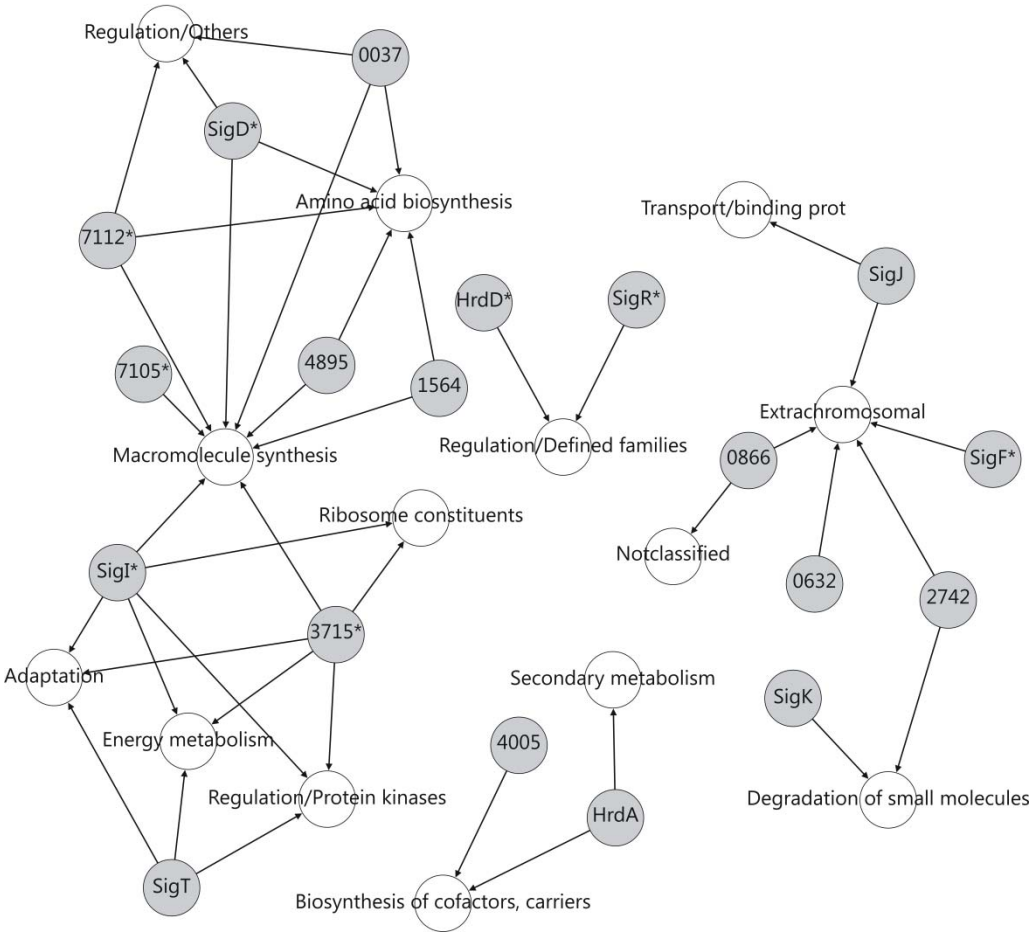


Figure 13. Suggested regulation of significantly enriched functional classes (white nodes) by sigma factors (gray nodes). Sigma factors are marked by name or gene SCO number. The asterisks designate sigma factors, whose overall expression was among 5% of the most expressed genes.

In the following investigation, we concentrated to two sigma factors, SigR and HrdD, that we suggested to be important for the progression of germination and we also proposed a cascade of events participating in recovery after breaking dormancy. Here, for the first time, we were able to suggest a role of particular sigma factors that act during germination in *Streptomyces*.

For the sigma factor SigR an advantage of combining our kinetic based approach with an independent binding data from ChIP-chip experiment (Kim et al. 2012) was taken. The ChIP-chip experiment predicted SigR static binding sites and thus reduced the number of SigR potential target genes to amount that could be computationally verified for possibility of kinetic relations. Sole binding of SigR to target gene promoter region is a necessary condition for transcription initiation, but it is not sufficient for proving functional relation (MacQuarrie et al. 2011; To and Vohradsky 2010). Therefore, a combination of kinetic expression data with static ChIP-chip binding suggestions represents useful method for identification of reliable regulator - target gene interactions.

From the proposed SigR target genes (Kim, et al. 2012), 145 were highly expressed in our dataset. Individual kinetic profiles of this set were tested by applying the kinetic transcriptional model (*1.14*) to determine whether, from the kinetic perspective, the suggested regulation is plausible. (The list of kinetically verified SigR target genes can be found in Supplementary Table 3 to Paper I provided on the attached CD).

Interestingly, the confirmed regulations implied that during germination, spores undergo similar response as under thiol-oxidative stress treatment (Kallifidas et al. 2010; Kang et al. 1999; Kim, et al. 2012; Paget et al. 1998; Paget et al. 2001; Park and Roe 2008). The similarity with thiol-oxidative stress response was supported by identification of related set of genes as under thiol-oxidative stress treatment (Kallifidas, et al. 2010; Kim, et al. 2012; Paget, et al. 2001). Among kinetically plausible SigR target genes were those involved in redox homeostasis (*rifO* - SCO7632, *trxA* - SCO3890, *trxB* - SCO3890, *trxC* - SCO0885, *mca* - SCO4967, *trxA4* - SCO1084, *msrA* - SCO4956, and *msrB* - SCO6061) and protein quality control (*clpP1* - SCO2619, *clpP2* - SCO2618, *clpC* - SCO3373, *clpX* - SCO2617, *prcA* - SCO1643, *mpa* - SCO1648, and *pepN* - SCO2643). Both genes engaged in redox homeostasis and protein quality control protect cellular compounds from irreversible damages that might result in protein misfolding and/or forming insoluble

aggregates. Revealed transcriptional control of proteins with general protective role brought important consequences in germinating spores; previous dormancy phase is characterized by increased presence of immobile protein aggregates (Cowan, et al. 2003; Mikulik, et al. 2002), that have to be either reactivated or metabolized, when damaged, during germination. Because in our experimental setup, SigR regulon expression was not induced by oxidative stress treatment, we propose that the probable stimuli of the stress that consequently induce SigR regulon expression in germination are provided by the high portion of aggregated proteins present, due to low water content, in the dormant spore.

This finding perfectly complements our observations based on proteomics experiments (Paper II, see also paragraph 3.3 below) where the expression of several chaperones and protein modifiers at the protein level was detected immediately after germination initiation. The function of chaperones and protein modifiers is, consistently with above claims, in assisting in protein folding (reactivation of present protein aggregates) and providing thus a sort of protein quality control in terms of aggregated and/or misfolded proteins.

The results also confirmed the global regulatory role of SigR as it controls expression of several regulatory proteins (*rsrA* - SCO5217, *ndgR* - SCO5552, *rsrA2* - SCO3451, *sigR2* - SCO3450, SCO1619, and SCO7140). Furthermore, we found that the SigR is a probable regulator of the alternative principal sigma factor HrdD.

The expression level of HrdD was detected approximately 20-fold higher than the average expression level in *S. coelicolor* germination dataset and the highest among all sigma factors. Kinetically possible target genes of HrdD were searched by applying the model *(1.14)* to all individual gene profiles. We found that most of the genes suggested to be under HrdD transcriptional control encod proteins that belong to defined regulatory families, such as transcription regulators of TetR, MarR, LysR, and GntR families. (The list of HrdD target genes can be found in Supplementary Table 4 to Paper I provided on the attached CD). In addition, we suggested that HrdD partially controls the expression of the chaperone-protease operon DnaK-HspR (*dnaJ* - SCO3669, *hspR* - SCO3668 and *clpB* - SCO3661). The operon also plays role in preventing proteins from forming aggregates or functionally or structurally damaged enzymes.

In summary, the above observations led to suggestion of successive events: re-hydration of the cell causes stress, similar to the thiol-oxidative stress. A probable

trigger of the stress that consequently induced SigR expression is a stimulus provided by high content of aggregated and/or misfolded proteins. The SigR regulates genes, whose products are involved in protein quality control and regulations. Among SigR target genes is also a gene for sigma factor HrdD that additionally controls expression of chaperones and other regulatory proteins that together participate on transition to active metabolism.

3.3 Biochemical processes associated with germination

To reveal metabolic processes associated with germination, proteomic-based approach in a time-dependent manner was employed. First, to describe accession of individual metabolic pathways, we focused on time when *de novo* synthesized proteins (radiolabeled) were firstly detected. The numbers of newly emerged proteins at specific time periods are depicted in Figure 14.

De novo synthesized proteins increments (Figure 14, the red parts of the columns) indicate that from T0 to T0.5 the spores are in a lag-phase. During this time, an elemental recovery after breaking dormancy occurs. Immediately after germination initiation (during the activation caused by heat shock treatment, T0) expression of 25 proteins started and other few newly synthesized proteins emerged within the next 30 min (T0.5). Among the proteins induced during spore activation, several chaperones and protein modifiers were detected (Tig – SCO2620, GrpE – SCO3670, DnaK – SCO3671, GroEL2 – SCO4296, GroEL1 – SCO4762, FkbP – SCO1638, CypH – SCO7510). The function of chaperones and protein modifiers is primarily in refolding and/or modification of aggregated proteins generated during sporulation. Reactivation of aggregated proteins is necessary to restart metabolic processes that utilize inner sources prepared during sporulation. To achieve this, the liquid influx into the cell is required, then the reactivation proceeds in assistance of chaperones and other enzyme modifiers that facilitates protein folding into their native forms.

From the germination initiation until the T1 stage, protein expression evokes the cell response to stress conditions. This claim was also supported by appearance of several regulatory proteins (BldG – SCO3549, Crp – SCO3571, Prs – SCO5244,

BldD SCO1489, CutR – SCO5862, RNase III – SCO5572, SCO2127, SCO3013, and SCO4232) at this stage.

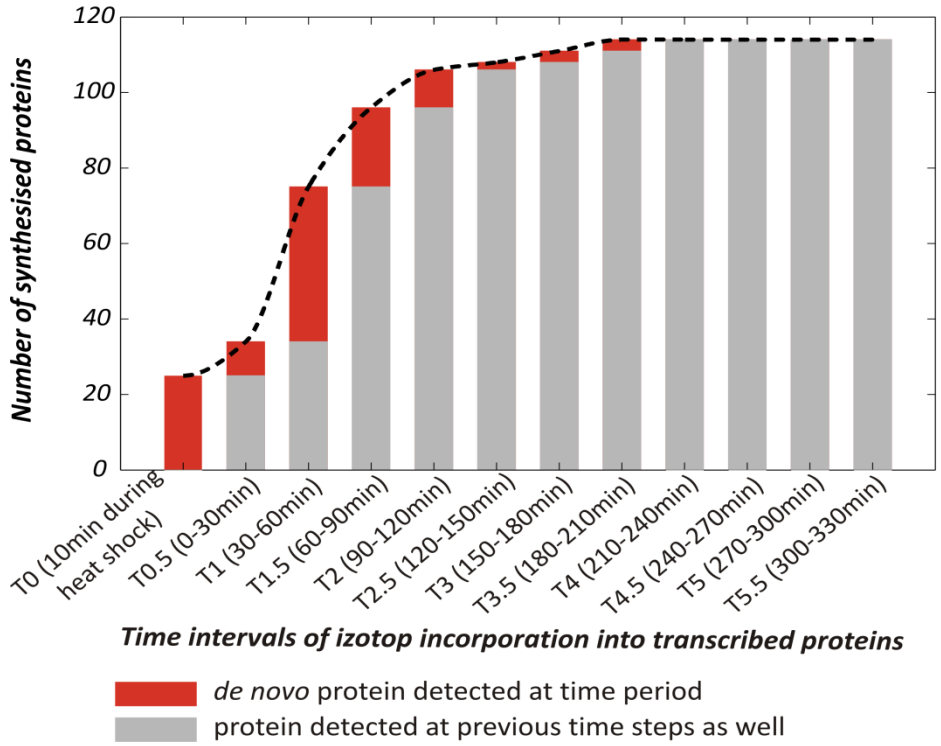


Figure 14. The columns show the numbers of synthesized (radioactive) proteins in measured time periods. The red part of the columns indicates the number of radiolabeled proteins whose expression is detected for the first time, while the gray part corresponds to the radiolabeled proteins whose expression continues from previous time steps.

The T1 period is characterized by rapid increase of proteosynthesis (Figure 14) when spores launch basal metabolism and are capable to detect environmental cues and respond to them by adjusting gene expression to actual requirements. Moreover, until the T1 stage, proteins important for developmental differentiation (DivIVA – SCO2077, FtsZ – SCO2082, Tdd8 – SCO2368, TerD – SCO0641) were expressed. Between T1 to T3.5 protein expression was being stabilized. Even though the synthesis of proteins continued to increase after the T1 stage, there was not detected any newly emerged proteins after the T3.5 stage.

The characteristics of individual protein functions are described in section "Analysis of Functional Groups" in Paper II. On the large-scale, our results imply that the most proteins necessary for germination start expression within 3.5 hours after germination initiation; the crucial period being up to 1 hour.

The initiation of expression of individual protein functional classes is shown in Figure 15, where only *de novo* synthesized protein increments at time periods are considered. The spores primarily activate expression of chaperones, translational machinery, and differentiation proteins that take control over energy metabolism and further development.

At the T1 stage, the peak of expression of regulators, transport/binding proteins and enzymes with unknown function was detected. Although, we have not studied the roles of these functionally uncharacterized proteins, their identification during germination might serve as a data source for further investigations and provide more comprehensive insight into the processes involved in control of germination. The list of individual identified proteins as well as assignment to the functional class and detected time of the first emergence are provided in Supplementary materials Table 3 to Paper II on the attached CD.

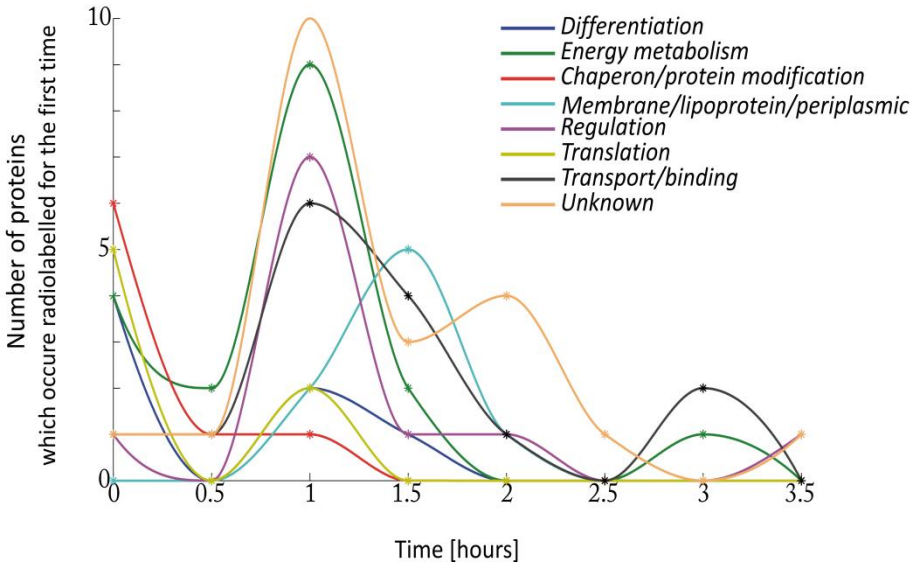


Figure 15. Progression in number of *de novo* synthesized proteins that initiate expression within the measured time periods for the 8 most occupied functional groups.

3.4 Comparative analysis

During *S. coelicolor* germination, large datasets of transcriptome expression time series (mRNA) and protein accumulation (Sypro stained) and protein synthesis (radiolabeling) time series were generated. Corresponding expression profiles were compared to investigate relationships between variables on different experimental stages.

3.4.1 Comparison of protein accumulation and synthesis

The dynamics of accumulated proteins (Sypro) reflects the rate of protein synthesis (radiolabeled) delayed by protein degradation and posttranslational modifications. To investigate the relationship between protein accumulation and protein synthesis, their profile kinetics were compared by determining Pearson correlation coefficient among the time series of the two experiments. The distribution of the correlation coefficients shows that the correlation between accumulation and synthesis is low or even negative (Figure 16) for approximately three-quarters of all compared profiles. The low correlation suggests differences in the rate of protein synthesis and accumulation, indicating that the effects of posttranslational modifications and degradation play a significant role in the control of gene expression during germination (Paper II).

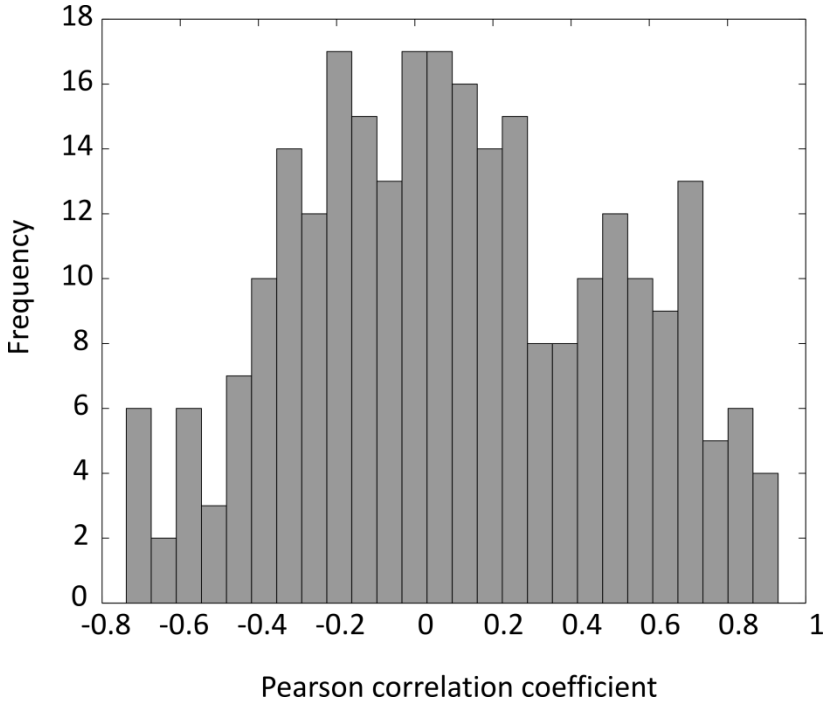


Figure 16. Distribution of Pearson correlation coefficients between Sypro stained and radiolabeled protein expression profiles.

3.4.2 Global features of transcriptome and proteome

More challenging is a comparison of transcriptome (mRNA) and proteome (Sypro and radiolabeled) expressions. There is a huge disproportion between transcriptome and proteome datasets because it is possible to measure thousands of genes on transcriptome level while only hundreds of features on protein level, giving the difference of at least one order in number of comparable profiles. Identified protein profiles are biased towards highly abundant proteins; therefore their selection is not a statistically representative sample of total protein pool. Thus, extrapolation of the results based on highly abundant proteins on the whole dataset should be taken with caution.

Here, instead of direct comparison of expression profiles, we focused on common features of the system that can be extracted from kinetic profiles. Kinetic profiles consisted of 13 measured time points, representing, in principle, a point in thirteen-dimensional space. Comparison of expression profiles means comparison of the distance between the points in the multidimensional space. Projecting optimally the points into lower dimensional space may reveal principal features of the

expression profiles by eliminating experimental and other noise. Such projection can be performed by Principal Component Analysis (PCA). Generally, few first PCs explain most of the variance of the dataset, higher PCs represent mostly experimental noise. In other words, PCA can extract and sort patterns of gene expression according to their contributions to the overall variance in the dataset from the most important to the most specific. We therefore projected the high-dimensional space onto a lower dimensional space by applying PCA both on transcriptomic and proteomic data and compared them on the first three principal components level (Paper III). Moreover, the principal component (PC) loadings shapes (temporal profile) refer to kinetics whose linear combination form individual kinetic profiles. Their sorting according to the amount of variance they bear, allows identifying principal kinetic patterns. Association of the PC patterns with individual gene kinetics can thus provide information about metabolic processes involved and determine those processes that are essential for progression of germination.

The PCA was applied individually for gene expression data (mRNA) and both protein expression data – accumulation (Sypro) and synthesis (radiolabeling). Trends of identified PC for all types of data are shown in Figure 17.

The correlation was found between three principal components: PC1 (mRNA) and PC2 (Sypro), PC2 (mRNA) and PC3 (Sypro), and PC3 (mRNA) and PC4 (Sypro), Figure 17 b, c, d. Similar trends in PCs for gene expression (mRNA) and protein accumulation (Sypro) were found, confirming that the fundamental processes are controlled in a coordinated fashion on both the transcriptomic and proteomic levels. On the contrary, protein synthesis (radiolabeling) was correlated only for PC1 (Figure 17b) but differed for higher PCs. In agreement with the above comparison (paragraph 3.4.1) of correlation coefficients of protein accumulation (Sypro) and synthesis (radiolabeling) – Paper II, the nature of these two processes seems to be different due to the involvement of other regulatory mechanisms. From that point of view, PCs from both protein accumulation (Sypro) and gene expression (mRNA) represent accumulative progression and thus similarity in PC loading trends was observed.

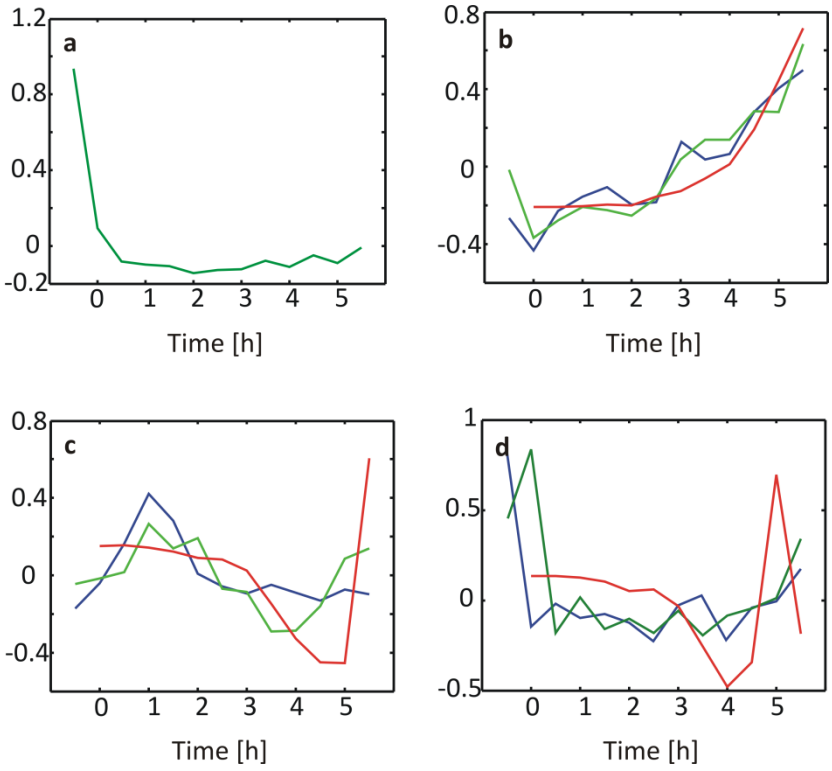


Figure 17. Profiles of the first PC loadings for the transcriptomic experiment and the two proteomic experiments. a) PC1 of the Sypro-stained proteomic experiment. b) blue – PC1 mRNA, green - PC2 Sypro, red – PC1 radiolabeling. c) blue – PC2 mRNA, green - PC3 Sypro, red –PC2 radiolabeling. d) blue – PC3 mRNA, green - PC4 Sypro, red – PC3 radiolabeling. The first time point in the radiolabeled profile is missing because this point represent dormant spores that were not radiolabeled.

From the assumption that PCs characterize principal information about the kinetics of the processes associated with PCs' trends (Alter et al. 2000; Alter et al. 2003; Holter et al. 2001; Holter et al. 2000; Vohradsky et al. 2007), we were able to identify the metabolic and regulatory pathways which are associated with the fundamental processes during *S. coelicolor* germination by computing correlation coefficients of these principal kinetic shapes (PCs) with individual expression profiles. The obtained groups of profiles correlated with PCs patterns were studied for characteristic gene functional assignments. For analyzing the correlation coefficients only gene expression data (mRNA) were used, because the numbers of identified proteins (Sypro and radiolabeling) in the functional groups were low for making global conclusions.

We found 1403 expression profiles associated with PC1 (Figure 17b, the blue curve) which represented continuously increasing trend in mRNA accumulation. The functional characteristics of PC1 associated genes showed that overrepresented functional classes of genes are mainly involved in basal metabolic processes (such as cell division, macromolecule synthesis/modification, ribosome constituents, energy metabolism), which are essential to activate after dormancy. 791 expression profiles were correlated with PC2 (a peak at T1 - Figure 17c, the blue curve). The PC2 associated genes represented the response of the cell to the actual environmental and / or inner requirements through enrichment of broad spectrum of genes with regulatory functions (functional classes regulation/defined families, regulation/others, transport/binding). At this point, based on transcriptome data analysis, we independently received consistent results with our proteomic analysis (Paper II and also paragraph 3.3 above), in which, at the T1 stage the synthesis of most of the regulatory and transport/binding proteins was detected in germinating spores. For the PC3 (Figure 17d, the blue curve), correlation was found for 342 expression profiles. The analysis of the overrepresented functional groups enabled to identify genes that reflected rapid switch of gene expression in amino acid biosynthesis and regulation/protein kinases functional classes within the first hour after germination initiation.

The criteria for identification of enriched functional groups are specified in Paper III. The genes assigned to the three PC loadings are listed in Supplementary Tables 2, 3 and 4 to Paper III on the attached CD.

After associating the major functional groups with PCs, we focused on the more specific classification of the genes essential for cell development (genes of energy metabolism, nucleic acids and protein synthesis) and also on the genes of the stress response, whose role interested us because, as we proposed in both Paper I and Paper II, germination can be considered as a reaction to the stress that is associated with the rehydration of the originally dry spores.

Generally, increment in abundance (correlation with PC1) was expectedly found for genes involved in the TCA cycle and its associated pathways because it is the backbone process of an active metabolism. The high correlation was also found for expression of genes whose products are involved in nucleic acids and protein synthesis, such as main elongation and translational initiation factors (e.g. *tsf* - SCO5625, *tuf* - SCO4662, *efp* - SCO1491, *fusA* - SCO4661, *infA* - SCO4725, and

infB - SCO5706). These results based on transcriptome (mRNA) analysis independently confirm our result from proteomic analysis (Paper II) that the re-activation of aggregated proteosynthetic components must be accompanied by the *de novo* synthesis of translational apparatus members. From the group of stress response genes, interestingly highly correlated with PC1 were almost all genes associated with cold shock (*scoF3* - SCO4684, *scoF* - SCO0527, *scoF2* - SCO4505, *scoF5* - SCO5921, *f40* - SCO3748, and *scoF1* - SCO3731). Because detailed role of individual cold shock proteins during developmental cycle is not known, we have no explanation about the role of cold chock proteins during germination and further research is required.

3.5 Authors' contributions to the study

Eva Straková (the author of the thesis) compared 2D gel electrophoresis images in PDQuest (Bio-Rad) software, analyzed measured data from microarrays and 2D gel electrophoresis (normalizations, statistics), applied kinetic model of transcription to time series expressions (computational and optimization procedures), interpreted results, and wrote Paper I, II and partially Paper III. Eva Straková also participated in wet lab experiments (bacteria cultivation, samples preparation, experimental methods optimization) mainly performed by technician **Alice Ziková**, who was in addition responsible for specialized techniques like manipulation with radioisotopes and large format 2D gel electrophoresis. **Jan Bobek** consulted experiments and results, helped with editing texts of Paper II and Paper III. **Pavel Řehulka** and **Helena Řehulková** performed mass spectrometry protein identifications; they created the Supporting Table 2 (to Paper II) and wrote a part of the Materials and Methods of Paper II related to the mass spectrometry. **Oldřich Benada** and **Olga Kofroňová** prepared electron microscopy images of spores in Paper I (Figure 1) and wrote a part of the Materials and Methods of Paper II corresponding to the electron microscopy. The whole project was designed and supervised by **Jiří Vohradský**.

4 Conclusions

In this thesis, the genes, proteins and processes related with *S. coelicolor* germination were characterized by employing computational systems biology approaches. The main results are summarized as follows:

- Within this work, concentrations of mRNA, accumulated and synthesized proteins during *S. coelicolor* germination were monitored at 13 time points by employing both DNA microarray chips and quantitative gel based proteomics. Within the large datasets, relatively densely sampled, kinetic profiles of 7115 genes, 671 Sypro stained proteins and 404 radiolabeled proteins were generated. In addition to the obtained results, the datasets can serve as data sources and reference materials for further investigations dealing both with germination and gene expression regulation in general. Currently, they represent the largest omics dataset not only in *Streptomyces* germination but in germination in general.

- The kinetic model of transcriptional regulation was applied to gene expression data, resulting in suggestion of hundreds of kinetically plausible target genes under individual sigma factors' control on a global scale. Computed parameters of the model for each individual interaction allow simulating kinetics of expression of controlled genes for different sigma factor expression profiles and consequently allow modeling of the kinetic behavior of the genetic network/networks. Particularly important for progression of germination, two sigma factors, SigR and HrdD, and their target genes were found. Another remarkable contribution of this study is a suggestion of regulatory interactions for previously unstudied sigma factors, whose functional characteristics should become a topic of individual studies.

- 251 proteins expressed during germination were identified by mass spectrometry. The subsequent inspection of proteosynthesis increments

revealed the advent of proteosynthesis mainly within the first hour of germination.

- The results imply that the primary concerns after hydration of spore cytoplasm and triggering germination include re-activation and functional qualification of proteins (mainly stored in a form of protein aggregates) present in the cell from the sporulation phase. The requirements of protein quality control were supported by both findings of regulatory interactions and observations of proteosynthesis. The specific genes and proteins participating in a protein protection were identified. The protein re-activation is accompanied by activation of main metabolic processes and *de novo* protein synthesis.

- Common features of the datasets were extracted by employing analysis of their principal components. Correlated expression profiles with the first three principal loadings were functionally analyzed allowing identification of metabolic processes (and individual genes) fundamental for the germination.

- The results on both transcript and protein levels imply that activation of metabolic processes immediately after breaking dormancy evokes cell response to stress conditions. This period is followed by rapid boost in expression of regulatory genes and proteins, approximately at one hour after germination initiation, indicating completion of assessment of the external sources to effectively adjust metabolism for further developmental growth through corresponding regulatory mechanisms.

5 References

AIJO, T. AND H. LAHDESMAKI Learning gene regulatory networks from gene expression measurements using non-parametric molecular kinetics. Bioinformatics, Nov 15 2009, 25(22), 2937-2944.

ALTER, O., P. O. BROWN AND D. BOTSTEIN Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci U S A, Aug 29 2000, 97(18), 10101-10106.

ALTER, O., P. O. BROWN AND D. BOTSTEIN Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. Proc Natl Acad Sci U S A, Mar 18 2003, 100(6), 3351-3356.

BEAL, M. J., F. FALCIANI, Z. GHAHRAMANI, C. RANGEL, et al. A Bayesian approach to reconstructing genetic regulatory networks with hidden factors. Bioinformatics, Feb 1 2005, 21(3), 349-356.

BENTLEY, S. D., K. F. CHATER, A. M. CERDENO-TARRAGA, G. L. CHALLIS, et al. Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). Nature, May 9 2002, 417(6885), 141-147.

BOBEK, J., P. HALADA, J. ANGELIS, J. VOHRADSKY, et al. Activation and expression of proteins during synchronous germination of aerial spores of Streptomyces granaticolor. Proteomics, Dec 2004, 4(12), 3864-3880.

BOSE, N. K. AND P. LIANG *Neural Network Fundamentals with Graphs, Algorithms, and Applications* McGraw-Hill Companies, 1996. ISBN 0070066183.

CLAESSEN, D., I. STOKROOS, H. J. DEELSTRA, N. A. PENNINGA, et al. The formation of the rodlet layer of streptomycetes is the result of the interplay between rodlins and chaplins. Mol Microbiol, Jul 2004, 53(2), 433-443.

COWAN, A. E., D. E. KOPPEL, B. SETLOW AND P. SETLOW A soluble protein is immobile in dormant spores of Bacillus subtilis but is mobile in germinated spores: implications for spore dormancy. Proc Natl Acad Sci U S A, Apr 1 2003, 100(7), 4209-4214.

DE CAMPOS, L. M. A scoring function for learning Bayesian networks based on mutual information and conditional independence tests. JOURNAL OF MACHINE LEARNING RESEARCH, 2006, 7, 2149-2187.

DEROUAUX, A., D. DEHARENG, E. LECOCQ, S. HALICI, et al. Crp of Streptomyces coelicolor is the third transcription factor of the large CRP-FNR superfamily able to bind cAMP. Biochem Biophys Res Commun, Dec 17 2004a, 325(3), 983-990.

DEROUAUX, A., S. HALICI, H. NOTHAFT, T. NEUTELINGS, et al. Deletion of a cyclic AMP receptor protein homologue diminishes germination and affects morphological development of Streptomyces coelicolor. J Bacteriol, Mar 2004b, 186(6), 1893-1897.

EDGAR, R., M. DOMRACHEV AND A. E. LASH Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res, Jan 1 2002, 30(1), 207-210.

FLARDH, K. Essential role of DivIVA in polar growth and morphogenesis in Streptomyces coelicolor A3(2). Mol Microbiol, Sep 2003a, 49(6), 1523-1536.

FLARDH, K. Growth polarity and cell division in Streptomyces. Curr Opin Microbiol, Dec 2003b, 6(6), 564-571.

GARDNER, T. S., D. DI BERNARDO, D. LORENZ AND J. J. COLLINS Inferring genetic networks and identifying compound mode of action via expression profiling. Science, Jul 4 2003, 301(5629), 102-105.

GHORBEL, S., A. SMIRNOV, H. CHOUAYEKH, B. SPERANDIO, et al. Regulation of ppk expression and in vivo function of Ppk in Streptomyces lividans TK24. J Bacteriol, Sep 2006, 188(17), 6269-6276.

GILLESPIE, D. T. The chemical Langevin equation. Journal of chemical physics, 2000, 113(1), 297-306.

GILLESPIE, T. D. Exact stochastic simulation of coupled chemical reactions. Journal of Physical Chemistry, 1977, 81(25), 2340–2361.

HAISER, H. J., M. R. YOUSEF AND M. A. ELLIOT Cell wall hydrolases affect germination, vegetative growth, and sporulation in Streptomyces coelicolor. J Bacteriol, Nov 2009, 191(21), 6501-6512.

HARDISON, R. C. AND J. TAYLOR Genomic approaches towards finding cis-regulatory modules in animals. Nat Rev Genet, Jul 2012, 13(7), 469-483.

HOLTER, N. S., A. MARITAN, M. CIEPLAK, N. V. FEDOROFF, et al. Dynamic modeling of gene expression data. Proc Natl Acad Sci U S A, 2001, 98(4), 1693-1698.

HOLTER, N. S., M. MITRA, A. MARITAN, M. CIEPLAK, et al. Fundamental patterns underlying gene expression profiles: simplicity from complexity. Proc Natl Acad Sci U S A, 2000, 97(15), 8409-8414.

KALLIFIDAS, D., D. THOMAS, P. DOUGHTY AND M. S. PAGET The sigmaR regulon of Streptomyces coelicolor A32 reveals a key role in protein quality control during disulphide stress. Microbiology, Jun 2010, 156(Pt 6), 1661-1672.

KANG, J. G., M. S. PAGET, Y. J. SEOK, M. Y. HAHN, et al. RsrA, an anti-sigma factor regulated by redox change. EMBO J, Aug 2 1999, 18(15), 4292-4298.

KIESER, T., M. J. BIBB, M. J. BUTTNER, K. F. CHATER, et al. *Practical Streptomyces genetics*. Norwich, UK, 2000. ISBN 0-7084-0623-8.

KIM, M. S., Y. S. DUFOUR, J. S. YOO, Y. B. CHO, et al. Conservation of thiol-oxidative stress responses regulated by SigR orthologues in actinomycetes. Mol Microbiol, Jul 2012, 85(2), 326-344.

LAWRENCE, N. D., M. GIROLAMI, M. RATTRAY AND G. SANGUINETT. Learning and Inference in Computational Systems Biology In.: The MIT Press, 2009.

LEE, T. I., N. J. RINALDI, F. ROBERT, D. T. ODOM, et al. Transcriptional regulatory networks in Saccharomyces cerevisiae. Science, Oct 25 2002, 298(5594), 799-804.

MACQUARRIE, K. L., A. P. FONG, R. H. MORSE AND S. J. TAPSCOTT Genome-wide transcription factor binding: beyond direct target regulation. Trends Genet, Apr 2011, 27(4), 141-148.

MATLAB. Product help. [R2010b]. The MathWorks Inc., 2010.

MIKULIK, K., J. BOBEK, S. BEZOUSKOVA, O. BENADA, et al. Expression of proteins and protein kinase activity during germination of aerial spores of Streptomyces granaticolor. Biochem Biophys Res Commun, Nov 29 2002, 299(2), 335-342.

NOENS, E. E., V. MERSINIAS, B. A. TRAAG, C. P. SMITH, et al. SsgA-like proteins determine the fate of peptidoglycan during sporulation of Streptomyces coelicolor. Mol Microbiol, Nov 2005, 58(4), 929-944.

PAGET, M. S., J. G. KANG, J. H. ROE AND M. J. BUTTNER sigmaR, an RNA polymerase sigma factor that modulates expression of the thioredoxin system in response to oxidative stress in Streptomyces coelicolor A3(2). EMBO J, Oct 1 1998, 17(19), 5776-5782.

PAGET, M. S., V. MOLLE, G. COHEN, Y. AHARONOWITZ, et al. Defining the disulphide stress response in Streptomyces coelicolor A3(2): identification of the sigmaR regulon. Mol Microbiol, Nov 2001, 42(4), 1007-1020.

PARK, J. H. AND J. H. ROE Mycothiol regulates and is regulated by a thiol-specific antisigma factor RsrA and sigma(R) in Streptomyces coelicolor. Mol Microbiol, May 2008, 68(4), 861-870.

PIETTE, A., A. DEROUAUX, P. GERKENS, E. E. NOENS, et al. From dormant to germinating spores of Streptomyces coelicolor A3(2): new perspectives from the crp null mutant. J Proteome Res, Sep-Oct 2005, 4(5), 1699-1708.

POLYNIKIS, A., S. J. HOGAN AND M. DI BERNARDO Comparing different ODE modelling approaches for gene regulatory networks. J Theor Biol, Dec 21 2009, 261(4), 511-530.

RANADE, N. AND L. C. VINING Accumulation of intracellular carbon reserves in relation to chloramphenicol biosynthesis by Streptomyces venezuelae. Can J Microbiol, Apr 1993, 39(4), 377-383.

RASMUSSEN, C. E. AND C. K. I. WILLIAMS *Gaussian Processes for Machine Learning*. the MIT Press, 2006. ISBN 026218253X.

SHMULEVICH, I., E. R. DOUGHERTY, S. KIM AND W. ZHANG Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. Bioinformatics, Feb 2002, 18(2), 261-274.

SHMULEVICH, I., H. LAHDESMAKI, E. R. DOUGHERTY, J. ASTOLA, et al. The role of certain Post classes in Boolean network models of genetic networks. Proc Natl Acad Sci U S A, Sep 16 2003, 100(19), 10734-10739.

SPALL, J. C. *Introduction to Stochastic Search and Optimization*. Wiley-Interscience, 2003. ISBN 978-0471330523.

SUSSTRUNK, U., J. PIDOUX, S. TAUBERT, A. ULLMANN, et al. Pleiotropic effects of cAMP on germination, antibiotic biosynthesis and morphological development in Streptomyces coelicolor. Mol Microbiol, Oct 1998, 30(1), 33-46.

TKACIK, G. AND A. M. WALCZAK Information transmission in genetic regulatory networks: a review. J Phys Condens Matter, Apr 20 2011, 23(15), 153102.

TO, C. C. AND J. VOHRADSKY Measurement variation determines the gene network topology reconstructed from experimental data: a case study of the yeast cyclin network. FASEB J, Sep 2010, 24(9), 3468-3478.

VAN BAKEL, H. AND F. C. P. HOLSTEGE A Tutorial for DNA Microarray Expression Profiling. Evaluating Techniques in Biomedical Research, CellPress Online, 2008, p.22-28.

VAN KAMPEN, N. G. *Stochastic Processes in Physics and Chemistry, 3rd edition*. North Holland, 2007. ISBN 978-0444529657.

VEITIA, R. A. A sigmoidal transcriptional response: cooperativity, synergy and dosage effects. Biol Rev Camb Philos Soc, Feb 2003, 78(1), 149-170.

VOHRADSKY, J. Neural network model of gene expression. FASEB J, Mar 2001, 15(3), 846-854.

VOHRADSKY, J., P. BRANNY AND C. J. THOMPSON Comparative analysis of gene expression on mRNA and protein level during development of Streptomyces cultures by using singular value decomposition. Proteomics, Nov 2007, 7(21), 3853-3866.

VU, T. T. Modeling and reconstruction of genetic networks. Universtiy of West Bohemia, 2005 doctoral thesis.

YU, J., V. A. SMITH, P. P. WANG, A. J. HARTEMINK, et al. Advances to Bayesian network inference for generating causal networks from observational biological data. Bioinformatics, Dec 12 2004, 20(18), 3594-3603.

6 List of abbreviations

2D	two-dimensional
cDNA	complementary DNA
GRN	gene regulatory network
ChIP	chromatin immunoprecipitation
PC	principal component
PCA	principal component analysis
SDS	sodium dodecyl sulphate
TF	transcription factor

7 Data on compact disc

The attached CD contains supplementary materials to papers listed in Appendices 8 and Figure 12 in EPS format.

Paper I (Appendix A)

Supplementary Table 1. The list of highly expressed sigma factors.

Supplementary Table 2. The list of genes assigned to the clusters.

Supplementary Table 3. Genes that represent a subset of SigR regulated genes.

Supplementary Table 4. The list of kinetically plausible HrdD target genes.

Paper II (Appendix B)

Supporting Information Table 1, normalized intensities of 2DE gel spots for both types of labeling;

Supporting Information Table 2, MS protein identification details;

Supporting Information Table 3, the list of identified proteins from germination of S. coelicolor and their ordering into functional groups and clusters;

Supporting Information Figure 1, the reference 2DE gel image of proteins from germination of S. coelicolor;

Supporting Information Figure 2, the reference 2DE gel image with marked identified proteins.

Paper III (Appendix C)

Supplementary Table 1 - Genes with highly correlated scores (Pearson correlation >=0.95) of Sypro Ruby-stained proteins (PC2) and mRNA (PC1).

Supplementary Table 2 - Genes that have an expression profile that is correlated with PC1.

Supplementary Table 3 - Genes that have an expression profile that is correlated with PC2.

Supplementary Table 4 - Genes that have an expression profile that is correlated with PC3.

Supplementary Figure 1 - Physiological change of spores during germination.

Supplementary Figure 2 – Comparison of the principal components with DNA synthesis.

8 Appendices

List of appendices:

Appendix A

Paper I

STRAKOVA, E., A. ZIKOVA and J. VOHRADSKY. Inference of Sigma Factor Controlled Networks by Using Numerical Modeling Applied to Microarray Time Series Data of the Germinating Prokaryote. *Manuscript in review*

Appendix B

Paper II

STRAKOVA, E., J. BOBEK, A. ZIKOVA, P. REHULKA, O. BENADA, H. REHULKOVA, O. KOFRONOVA and J. VOHRADSKY. Systems Insight into the Spore Germination of *Streptomyces coelicolor*. *J Proteome Res*, Jan 4 2013, 12(1), 525-536.

Appendix C

Paper III

STRAKOVA, E., J. BOBEK, A. ZIKOVA and J. VOHRADSKY. Global Features of Gene Expression on the Proteome and Transcriptome Levels in *S. coelicolor* during Germination. PLoS ONE, *accepted July 2013*

Appendix A

Paper I

Inference of Sigma Factor Controlled Networks by Using Numerical Modeling Applied to Microarray Time Series Data of the Germinating Prokaryote

Strakova, E., Zikova, A. and Vohradsky, J.

Manuscript in review

INFERENCE OF SIGMA FACTOR CONTROLLED NETWORKS BY USING NUMERICAL MODELING APPLIED TO MICROARRAY TIME SERIES DATA OF THE GERMINATING PROKARYOTE.

EVA STRAKOVA[1], ALICE ZIKOVA[1] AND JIRI VOHRADSKY[1,*]

[1] Laboratory of Bioinformatics, Institute of Microbiology, Academy of Sciences of the Czech Republic, Prague, 142 20, Czech Republic

*To whom correspondence should be addressed.
Tel: +420241062513
E-mail: vohr@biomed.cas.cz

ABSTRACT

A computational model of gene expression was applied to a novel test set of microarray time series measurements to reveal regulatory interactions between transcriptional regulators represented by 45 sigma factors, and the genes expressed during germination of a prokaryote *S. coelicolor*. Using microarrays, the first 5.5 hours of the process was recorded in 13 time points, which provided a database of gene expression time series on genome-wide scale. The computational modeling of the kinetic relations between the sigma factors, individual genes and genes clustered according to the similarity of their expression kinetics identified kinetically plausible sigma factor controlled networks. Using genome sequence annotations, functional groups of genes that were predominantly controlled by specific sigma factors were identified. Using external binding data complementing the modeling approach, specific genes involved in the control of the studied process were identified and their function suggested.

1 INTRODUCTION

The expansion of high-throughput techniques in recent years has increased the potential to infer new biological knowledge from existing data and has also increased the demands of computational approaches to decipher large quantitative datasets. One of the primary challenges of systems computational biology lies in inferring gene regulatory networks among genes from time series expression data. A typical source of genome-wide information represent gene expression data obtained from microarrays that can be used for inference of transcriptional control networks. In this paper, we focused on identifying the potential target genes of transcription regulators using a reverse-engineering transcriptional model based on the relationship between regulator expression profiles and the expression of its target genes.

Numerous computational approaches have been employed to identify regulatory interactions between genes, including ordinary and stochastic differential equations, neural networks, dynamic Bayesian networks, and information theoretic- or correlation-based methods, which are reviewed in papers of Bansal et al. or Penfold and Wild (1,2). Generally, the methods differ in prior knowledge requirements and experimental data.

The majority of transcriptional models (also the one used in this paper) are based on the assumption that the dynamics of the regulator at the protein level is correlated with the dynamics at the transcript level (3); therefore, measured protein levels can be replaced with relatively straight forwardly measured transcriptome profiles. The main reason for the approximation lies in the fact that protein dynamics currently cannot be easily measured on a global scale. The approach of Gao et. al. (4) addresses latent regulator concentration by employing Gaussian process inference techniques into the transcription model. The Gaussian process together with the regulator translation model were then used to rank target genes on a genome-wide scale to identify potential target genes (5), and further extension of the single regulator model was conducted by involving multiple interacting regulators (6).

In bacteria, the initiation of transcription depends mainly on sigma factors, which are proteins (regulators) that are able to recognize and bind, in the form of a RNA polymerase holoenzyme, to a specific gene promoter region (target gene) and guide RNA polymerase to start transcription. Therefore, sigma factors are the essential nodes in gene regulatory networks that govern further interactions and

processes in the cell. A crucial task involved in inferring gene regulatory networks in bacteria is the recognition of the target genes of sigma factors. Experimentally, the physical interaction between sigma factors and gene promoter sequences is verified by chromatin immunoprecipitation methods (ChIP-chip and ChIP-Seq). It was shown, however, that the static binding information may also include silent binding events that do not directly enhance transcription (7,8). A combination of kinetic expression data with static binding site predictions represents an advantageous approach for inferring functionally related interactions between sigma factors and target genes. The dynamic model of gene expression was used here to explore the kinetically plausible regulatory relationships between sigma factors and their potential target genes based on the newly generated time series dataset mapping the germination of *Streptomyces coelicolor*.

Streptomyces species are Gram-positive soil bacteria that are widely studied for two primary reasons. First, they are important natural producers of diverse antibiotics and biologically active compounds. Second, due to their complex developmental lifecycle (including single spore germination followed by vegetative mycelia formation, aerial hyphae growth and unigenomic spore formation), *Streptomyces* serve as model organisms for fundamental cell development studies.

During dormancy, spore content is protected against unfavorable conditions by a complex coat structure, and the metabolic activity of the cell is minimal in this life-phase. The process of breaking dormancy and awakening the cell to an active metabolism is called germination. The regulation of germination is important, however, poorly understood area of *Streptomyces* biology (9-12). For systems studies, the transition from dormancy to vegetative growth represents an excellent model process due to a well-defined initial state, when the development of the system always begins from the consistent pool of protein and RNA molecules.

Individual life-cycle stages are characterized by both different metabolisms and physiologies as well as by the involvement of different regulatory and signaling pathways. Numerous proteins with regulatory functions (approximately 12%) are predicted to exist in the *S. coelicolor* genome (13). Recent studies have also suggested an important role of regulatory RNA molecules (14,15). The *S. coelicolor* genome possesses 65 annotated sigma factors (13). Thus, in comparison with other bacteria such as *Mycoplasma genitalium* (1 sigma factor), *Escherichia coli* (7 sigma factors) or *Bacillus subtilis* (18 sigma factors), *Streptomyces* has an enormous

capacity for regulation. The complexity of regulation in *Streptomyces* has fascinated researchers for decades. Current systems approaches applied on a global genomic scale, such as transcriptomics and chromatin immunoprecipitation methods (16-20), contribute to the unraveling of this regulatory complexity.

Only a small number of *S. coelicolor* sigma factors have been functionally characterized. For example, the principal sigma factor HrdB (SCO5820) represents the primary housekeeping regulator. Similar to the primary sigma factor and also closely related in promoter recognition are three sigma factors, HrdA (SCO2465), HrdC (SCO0895) and HrdD (SCO3202); however, these three factors have been reported to be non-essential for exponential growth (21). SigB (SCO0600) and SigH (SCO5243) play important roles in the osmotic stress response (22), whereas SigH has been also suggested to influence morphological differentiation (23). SigK (SCO6520) appears to negatively control development and antibiotic production (24). Other developmental sigma factors involved in differentiation are SigF (SCO4035), the late sporulation gene that affects spore maturation (25); as well as SigN, which is believed to control aerial hyphae composition (26); WhiG (SCO5621), which is involved in sporulation by initiating the *whi* gene cascade (27); and BldN (SCO3323), which has been suggested to participate in sporulation control (28). SigT (SCO3892) may negatively influence differentiation and secondary metabolism (29). SigE (SCO3356) was suggested to be an important regulator of cell wall biosynthesis (30). The sigma factor SigR (SCO5216) was studied for its cell defense role against thiol-oxidative stress (31-34) and protein quality control (20,35).

Most of the mentioned functional characteristics were obtained by observing mutant phenotypes and expression under different experimental conditions; however, these methods do not usually identify molecular mechanisms or direct interactions between sigma factors and their target genes. Several studies have focused on identifying potential target genes of sigma factors using both in vitro and in vivo experiments and by examining the interaction between sigma factors and a vast variety of anti-sigma factors and anti-anti sigma factors (20,33,35-37).

In this study, we applied a numerical model of gene expression kinetics to identify potential sigma factor target genes in *S. coelicolor* wild-type expression data. The used model (38) originating from formalized recurrent neural networks was derived under the consideration that transcription is a temporal dynamic action and can be described using a system of differential equations. Further evaluation of the

model parameters led to the computation of the expression profiles of target genes that were then compared with measured microarray data. We monitored quantitative changes in the transcriptome over time during *S. coelicolor* germination, generating thus a large experimental dataset. We measured dynamic changes in the transcriptome at 13 time points during the initial 5.5 h of *S. coelicolor* germination using microarrays (37 microarrays in total). The resulting relationships between sigma factors and regulated genes or groups of genes were interpreted in biological manner and compared with published data.

2 MATERIALS AND METHODS

2.1 Cultivation and germination

The details regarding *S. coelicolor* A3(2) M145 spore cultivation and growth were published in our previous work (12). Briefly, spores harvested from agar plates (growth for 14 days) were germinated in liquid AM media at 37°C. Spores were activated by mechanical disruption of the outer coat, and a 10 min heat shock treatment was applied to boost synchrony. Samples for RNA isolation were collected during 5.5 hours of germination in 30 min time intervals. Altogether, we obtained samples at 13 time points, including samples from dormant spores.

2.2 RNA isolation from spores

To break the cells, we used a FastPrep-24 machine (Biomedicals) where the spores were mechanically disrupted in tubes containing zirconium sand, two 4-mm glass beads, 500µl of lysis buffer (39) (50mM Tris-HCl pH8, 500mM LiCl, 50mM EDTA pH8, and 5% SDS) and 8µl of RNAse inhibitors (Biorad).The disruption occurred in 6 rounds for 35 s while the tubes were re-chilled between each round. The samples were centrifuged at 14000 g for 15 min at 4°C, and phenol-chloroform RNA extractions were performed on the supernatant twice. The RNA was precipitated overnight in ethanol and 3 M sodium acetate at -20°C. Finally, the RNA was re-suspended in 50µl RNAse-free water and 0.5 µl RNAse inhibitors, and the remaining DNA was removed using a DNAse Free kit (Ambion). The RNA was stored in water at -20°C.

2.3 DNA microarrays and data processing

RNA quality control and gene expression levels were performed by Oxford Gene Technology (Oxford, UK) using Agilent DNA microarrays covering the entire *S. coelicolor* genome and the standard Bacterial RNA amplification Protocol for two-channel assays by OGT.

The acquired data were linear LOWESS normalized and filtered for background and flag information (from Agilent documentation) in the GeneSpring software to obtain genes that were significantly expressed above background and to avoid side effects of possible cross hybridizations. These methods reduced the number of entities on a single array from 43888 to 25312, which finally represented the outcome for 7115 genes out of 7825. The data discussed in this publication have been deposited in the NCBI Gene Expression Omnibus (40) and are accessible using the GEO Series accession number GSE44415 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE 44415).

2.3.1 Array normalization

The experiment included 37 arrays from 13 distinct time points of *S. coelicolor* germination. The arrays shared a common reference in the red channel (Cy5), which was a mixture of RNA samples from all examined time points. The distributions of Log2Ratio values (Log2Ratio = log2 (Sample (Cy3)/Reference (Cy5))) from each array were centered to ensure that the medians and the median absolute deviations of all the array distributions were equal. The centering was performed by subtracting the Log2Ratio median value of the array from each Log2Ratio measurement on the array and was divided by the median absolute deviation. To eliminate array outliers, we filtered out the 0.02 quantile of the least and the most intensive Log2Ratio values. Therefore, normalized Log2Ratios were exponentiated to return the values to the original scale (normalized Ratios)

2.3.2 Gene expression profile inference

Time series of relative mRNA concentrations (“gene expression profiles”) were obtained by averaging the normalized Ratios across biological replicates at specific time points and across all the gene replicate spots on the array. Before averaging, the outliers among gene replicates at individual time points were filtered using the Q-test (for 3-9 inputs) and the Pierce test (for > 10 inputs).

The filtering caused a few profiles to have no value for certain time points (= 0). These zero values were examined to determine if they occurred between two non-zero time points. In this case, the neighboring time points were non-zero; therefore, the missing value was linearly interpolated (performed for approximately 100 profiles out of 7115).

2.4 Expression profile analysis

2.4.1 Highly expressed genes

To eliminate profiles with overall low expression during germination, we analyzed microarray sample channel signals (Cy3 labeling). The idea was to minimize the influence of gene profiles whose microarray signal originated from experimental errors that exceeded the pure technical limits for eliminating signals under the background. Thus, for each gene, the overall expression level was specified by computing the median across all microarray replicates at all time points for the sample channel microarray signal. The profiles, whose overall expression level was below the first quartile value (563) of all counted medians, were filtered (category I. in Figure 1A). To avoid omitting profiles with a low overall expression level but with a significant peak, the filtered expression profiles were further confirmed. In the presence of a significant peak, the profile was considered as highly expressed and added to the set. The final set of "highly expressed" genes contained 5385 profiles and was used in further analysis.

2.4.2 The selection of genes with small variance

To reduce the influence of genes with high measurement error on core profile inferring, the coefficients of variation (CV) were computed for all genes at each time point. From the entire set of genes, only profiles that met the criteria of having a CV of less than 0.47 for at least 8 time points (out of 13) in the profile were chosen. The resulting set of genes contained 3317 genes, which were further used to infer the kinetic cluster core profile.

2.4.3 Core profiles

The set of genes with a relatively small CV was grouped according to similarity in their expression profiles. The grouping was conducted in the same manner as in

our previous work (12). To determine a typical kinetic profile of the particular group, we employed the k-means clustering method with a Spearman correlation as a distance metric. The k-means algorithm was repeated 500 times for a predefined number of clusters (n). For each cluster, we then defined a cluster core as a group of profiles that appeared in one cluster in at least 50% of the repeated runs.

To roughly estimate the optimal number of clusters (n), the clustering procedure was re-computed for different numbers of clusters (n=30-70), and a jackknife method was also applied. The jackknife was based on the systematic re-computation of clusters while omitting a small amount of random observations (1.5%). The within-cluster sums of point-to-centroid distances (J) (41) of each resulting core cluster were assessed and plotted against the number of clusters (n) (data not shown). The point on the curve where the significant local change of J occurred indicated the potentially optimal n for the given dataset. The optimal number of clusters for the dataset was n=42.

For all 42 kinetic core clusters, the average profile (core profile) of each core cluster was calculated. The core profiles were then used as inputs for the modeling procedures and as the seeds for the classification of profiles into 42 groups according to their correlation with the core profile.

2.5 Model of gene expression

We employed the kinetic model of gene expression suggested by Vohradsky (38), which was later revised and extended (42,43). The model is derived from the assumption that the actual concentration of a regulator (sigma factor in this study) determines the expression kinetics of a transcribed gene. Generally, the possibility of triggering gene transcription depends on the likelihood of sigma factor binding (conjugated in RNA polymerase holoenzyme) to the promoter region. The binding probability of the sigma factor is determined by the “binding strength” of the specific sigma factor to the specific gene promoter and the number of sigma factor molecules around the DNA, which is proportional to the total amount of sigma factors. When the number of regulator molecules is low, the overall regulatory effect on gene expression is small. When the number of regulator molecules increases to a certain level, the transcription rate increases and the regulatory effect on the gene expression rate is proportionally modulated by the amount of regulator molecules. The process alters until the promoter becomes saturated and an increasing number of regulator

molecules do not increase the expression rate. The corresponding mathematical terms of the described relationship between the sigma factor concentration and gene expression rates represent a sigmoid function. Under previous considerations, the model for transcription control has the following form:

$$\frac{dy_i}{dt} = \frac{k_{1i}}{1+\exp\left[-(w_i R_j + b_i)\right]} - k_{2i} y_i, \qquad \text{(Eq. 1)}$$

where y_i represents the transcript concentration of the regulated *i-th* gene, and R_j is the transcript concentration of the *j-th* sigma factor (regulator), which is modulated by parameter w_i. The b_i parameter corresponds to a reaction delay. Incremental expression is diminished by the rate of transcript degradation described by the term $k_{2i}y_i$. The k_{1i}, k_{2i}, b_i and w_i model parameters are derived from experimental expression data, specifically microarray data in this study.

2.5.1 **Parameter optimization**

To fit measured gene expression time series y_{m_i} by the function y_i, given by (*Eq. 1)*, we used an optimization procedure and computed the k_{1i}, k_{2i}, b_i and w_i parameters of the model *(Eq. 1).* For each gene the parameters of the model (k_{1i}, k_{2i}, b_i and w_i) were optimized to fit the measured expression profile $y_{m_i}$of the *i-th* gene using the measured expression profile of the *j-th* sigma factor R_j by minimizing an objective function

$$F_i(k_{1i}, k_{2i}, b_i, w_i) = 1 - c_i(k_{1i}, k_{2i}, b_i, w_i), \qquad \text{(Eq. 2)}$$

where c_i represents a Pearson correlation coefficient calculated for pair y_{m_i} (measured expression profile) and $y_i(t,R,k,b,w)$ (computed expression profile).

Simulated annealing (44) was used as the minimization procedure. Simulated annealing performs well when the parameter space contains more local minima in which other optimization procedures can be trapped, which was considered for our data. The resulting parameters k_{1i}, k_{2i} and w_i, were forced to remain positive to reflect their biological nature.

Equation 1 was solved numerically. For numerical evaluation, the Matlab function ode45 based on an explicit Runge-Kutta formula was used.

Initial parameters were pre-set by random values, and then the optimization procedure was performed. For each examined relationship between sigma factor and target gene or sigma factor and another sigma factor or sigma factor and typical representative profile of kinetic cluster (core profile), the optimization procedure was completed 15 times with diverse initial parameter values. The optimal set of parameters was established as the set with the lowest value of the objective function (*Eq. 2*); therefore, the highest correlation occurred between the modeled and measured expression profiles.

The criterion for the goodness of fit between the measured profile y_{m_i} and the modeled expression curve y_i was the Pearson correlation coefficient. When the correlation coefficient exceeded a predefined value, the interaction between the sigma factor and a gene was considered possible. For the modeling of regulations where no prior knowledge was available (paragraphs 3.4.1, 3.4.2 and 3.5.2), the Pearson correlation coefficient was required to be higher than 0.8. For the modeling of interactions found by the ChIP-chip experiment or in the literature (paragraphs 3.5.1 and 3.5.3), the requirement for the Pearson coefficient was arbitrarily set to 0.65 to obtain all possible and even less correlated interactions. The interactions that satisfied the criteria were further visually validated.

2.6 Visualization of networks

The open source software Gephi https://gephi.org/ was used for network visualization.

3 RESULTS AND DISCUSSION

3.1 Experimental design

The *Streptomyces coelicolor* spores evolved during the examined period of 5.5 h from the dormant stage to cells with germ tubes.

To understand transcriptional regulations in germination in *S. coelicolor*, we employed a transcriptomic-based approach in a time-dependent manner. During the monitored 5.5 h, the RNA samples were collected at 30 min time intervals. The RNA sample collected from dormant spores was set as the initial time point (T Dorm), followed by RNA sample obtained after heat shock treatment (T0) and continued by samples T0.5 – T5.5 gained in 30 min intervals. Finally, we obtained RNA samples

from 13 time points. For each of the 13 time points, RNA was isolated from three, for time points 4 and 7 from two, independent cultures. The mRNA expression levels were measured by microarray. In total, entire experimental set contained 37 microarrays.

3.2 Highly expressed sigma factors in germination

The term "gene expression profile" used through the text refers to the normalized Ratio signals as described in the Methods paragraph 2.3.2, which recorded temporal changes in mRNA expression kinetics. The arrangement of the microarray experiment (sample mRNA –Cy3 channel; reference - mixture of total mRNA – Cy5 channel) enhanced measurement accuracy but did not provide information regarding the absolute expression levels of individual genes due to the various hybridization levels of the reference caused by the diverse probes on the microarray. If we consider that an equal amount of mRNA was always loaded onto the microarray chip, we only used the sample channel signals (Cy3) to estimate the absolute expression levels of individual genes. Although this approach led to increased variance of the averaged expression values, the expression kinetics of the sample channel and kinetics of the normalized Ratios stayed highly correlated (data not shown), indicating that the overall kinetic trends were similar for both types of data. The overall expression levels based on the sample channel signals were calculated (Methods 2.4.1) for each gene. The logarithmic distribution (based 2) of the overall expression levels was approximately lognormal (Figure 1A), with the long right tail representing approximately 5% of genes with extreme overall expression in comparison with the entire dataset.

Among the highly expressed genes (overall expression level above the first quartile, categories II. – IV. Figure 1A), 45 of the 65 annotated sigma factors in the *S. coelicolor* genome were detected. The overall expression levels of the most transcribed sigma factors (category IV. Figure 1A) are shown in Figure 1B, and all highly expressed sigma factors are listed in Supplementary Table 1. This paper discusses the identification of regulatory interactions of these highly expressed sigma factors with individual genes and gene kinetic clusters.

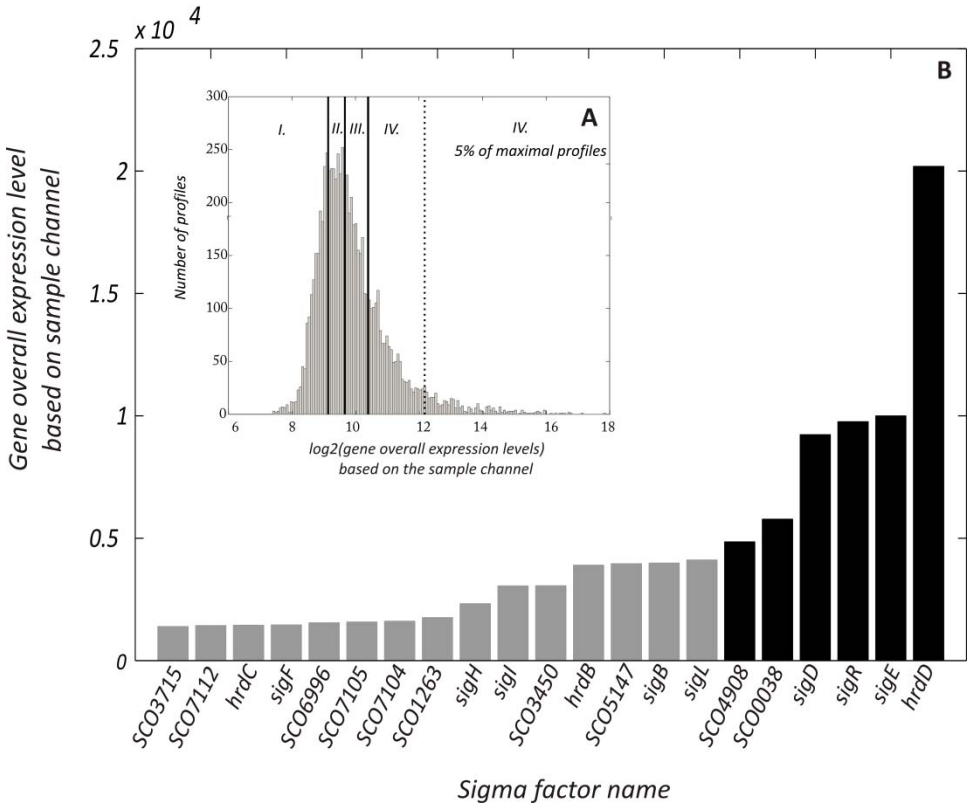


Figure 1. A) The distribution of log2 gene overall expression levels. The black vertical lines indicate 1st, 2nd and 3rd quartiles; the dotted line indicates 5% of the most expressed genes. The highly expressed set includes genes above the 1st quartile (categories II. – IV.; 75% of the dataset). B) The overall expression level of the 21 most highly expressed sigma factors during *S. coelicolor* germination. The grey columns represent genes with overall expression above the 3rd quartile (category IV.), and the black columns represent genes included in 5% of the most expressed genes.

3.3 Modeling of sigma factor transcriptional control

A crucial step for transcription initiation in bacteria is the recognition and binding of the sigma factor to the gene promoter region, enabling RNA polymerase to transcribe the gene into mRNA. The *S. coelicolor* genome possesses 65 annotated sigma factors that form a broad range of sigma factor–regulated gene combinations. Cell selection of an expression program, which is governed by specific sigma factors and signaling pathways, depends on the developmental stage of the cell and external

conditions. The primary task was to identify sigma factors significant for directing particular developmental processes and also identify their target genes.

For estimation of such regulatory effects we used a kinetic transcriptional model *Eq.1*. In principle, the model tests whether for the measured expression profile of the regulator exists a set of parameters that is able to simulate the expression profile of a target gene that would fit, with good accuracy, the expression profile that was measured. The genes with a good fit between their measured expression profiles and modeled expression profiles were indicated as kinetically possible target genes of sigma factors.

The investigation of one-by-one gene regulatory interactions in the entire dataset is extensively computationally demanding if all sigma factor–gene combinations have to be inspected (for 45 sigma factors and 7115 expressed genes it is 320 175 combinations). In reality, genes controlled in the same way share the same kinetic profile pattern. Instead of computing one to one interactions it is therefore possible to compute interactions between sigma factors and characteristic gene profiles based on the kinetic profile common for a group of genes without loose of generality. With this assumption in mind we identified kinetic clusters of genes having common expression profile and modeled the interactions on global scale between all 45 sigma factors and characteristic kinetic profiles of the clusters (paragraph 3.4)

To gain more detailed insight, the individual one-by-one strategy was applied solely to identify target genes of sigma factors HrdD (paragraph 3.5.2) and SigR (paragraph 3.5.1) which were selected for the following reasons - HrdD represented the most expressed sigma factor in the experiment and for SigR we were able to incorporate static ChIP-chip binding data (20). Individually were also examined interaction between sigma factors and their target genes that were proposed in literature (paragraph 3.5.3).

3.4 Global kinetic analysis of time series

3.4.1 Regulatory interactions between sigma factors and groups/clusters of similarly expressed genes

To compute all the potential regulatory combinations of the 45 highly expressed sigma factors and 7115 expressed genes, we would have to analyze 320 175 sigma factor target gene combinations. Therefore, keeping in mind that genes controlled in the same way have the same expression profile, instead of investigating one-by-one

gene regulatory interactions, we analyzed the transcriptome on a global scale by working with typical kinetic trends characteristic for group/cluster of genes with "similar" expression profiles that, from the kinetic point of view, may be regulated in the same way.

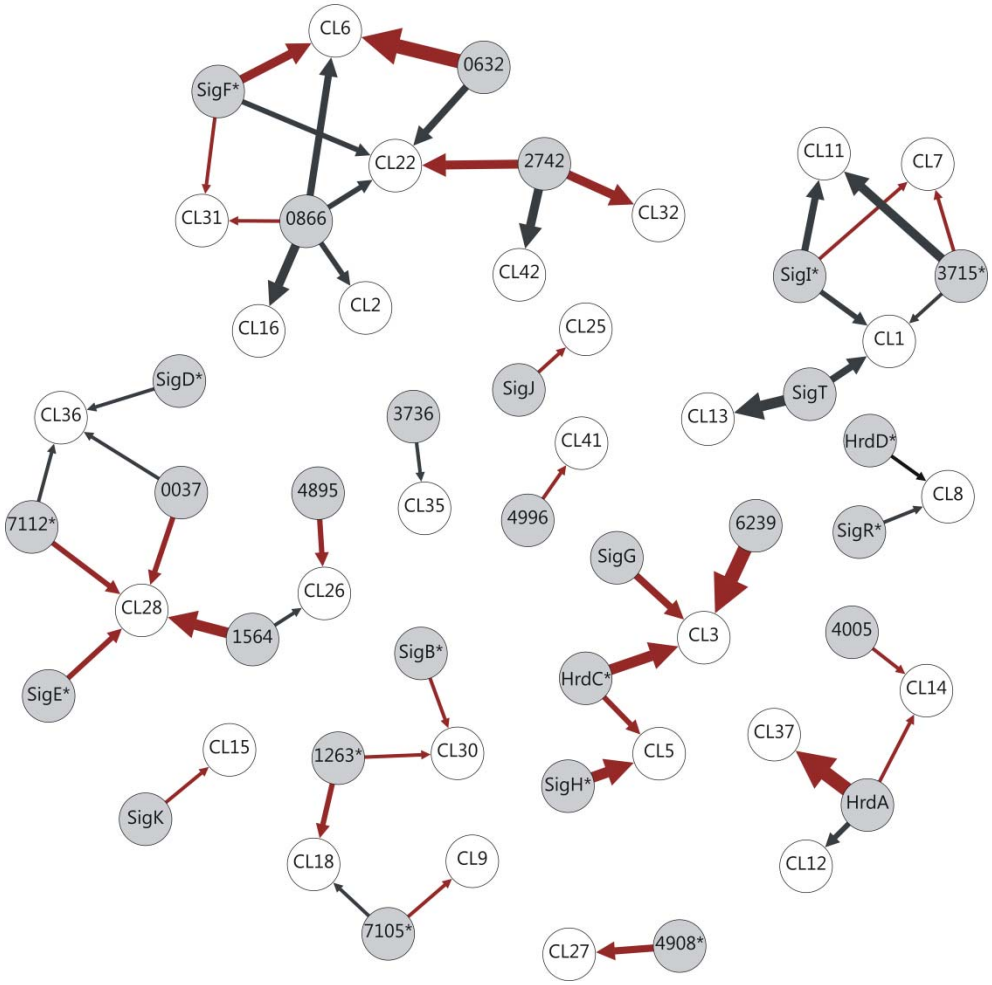


Figure 2. Suggested transcriptional regulation between the sigma factors (grey nodes) and the clusters/groups of genes with similar profiles (white nodes). The black arrows mark trivial regulations, while red arrows mark nontrivial regulations. The thickness of the arrows is proportional to parameter w of the model and corresponds to the strength of the regulatory effect. Sigma factors are marked by name or SCO number; kinetic clusters are marked with cluster number and their gene members are listed in Supplementary Table 2. Asterisks designate high overall expression of sigma factors from category IV (Figure 1A).

Each typical trend (defined by the core profile of the kinetic cluster – Methods 2.4.3) was tested as a possible target expression profile of each of the 45 studied

sigma factors. To identify the typical kinetic representatives of the target genes, we first selected a subset of gene profiles with low coefficients of variation within the experimental repeats to eliminate the influence of the profiles with higher measurement errors (Methods 2.4.1). Among the selected subset of 3317 genes with low CV, 42 different kinetic groups were identified. The core profiles were determined as an average profile of the most frequently occurring members in the particular group in the repeated runs of clustering (Methods 2.4.3).

For each pair sigma factor–core profile, the kinetic model (*Eq.1.*) and the optimization procedure (*Eq.2.*) were performed to identify kinetically possible regulations. Specifically, by assigning a kinetic cluster under sigma factor transcriptional control, we assumed that all members of the cluster were controlled by this sigma factor. Obtained results are referred in Table 1, the second column.

Figure 2 presents a visual representation of the computed interactions between the sigma factors (grey nodes) and typical representative profiles of the kinetic clusters (white nodes). The thickness of an arrow corresponds to the level of influence of the regulator to the target gene and is proportional to parameter *w* of the model (*Eq.1*). Kinetically possible regulation of a group/cluster of similarly expressed genes was found for 29 sigma factors. Remaining 16 sigma factors did not show any interactions that would satisfy the goodness of fit criteria.

A unique case occurred when the expression profile of the regulator and its target profile were highly correlated. Based on the principles of the model (*Eq. 1.*), these regulatory interactions represented “trivial” regulation. There are two possible interpretations of the trivial regulations. First, the response of the target gene to the regulator was very fast; therefore, their expression profiles had a similar course. Second, the regulator and the controlled genes had a common regulator that controlled them in a similar manner. Both possibilities are equally probable, and the method used here is unable to distinguish between them. In Figure 2, the trivial regulations are marked with black arrows, while the other regulations are marked in red.

When more than one proposed regulatory interaction exists (more arrows are pointing to one kinetic cluster in Figure 2), it should be emphasized that all alternatives are equally probable based on the kinetic aspects and cannot be interpreted as simultaneous regulations by multiple sigma factors. Figure 2 shows all

kinetically possible alternatives of regulation, and the gene members of individual kinetic clusters are listed in Supplementary Table 2.

All alternative principal sigma factors (HrdA, HrdC and HrdD), whose functions are still unknown, were suggested to control gene kinetic clusters 12, 37 (HrdA), 3 (HrdC) and 8 (HrdD). In addition to several known sigma factors (SigB, SigF, SigG, SigH, SigI, and SigK) and extracytoplasmic function subfamily (ECF) sigma factors (SigE, SigT, SigD), we proposed possible regulatory activity for many other sigma factors whose functions have not been previously identified (marked by SCO number in Figure 2).

3.4.2 **Regulatory interactions between regulators**

The same transcriptional regulatory mechanism controlling sigma factor–target gene relationships also occurs for sigma factors themselves. Thus, sigma factor expression is regulated by the interaction of either different sigma factors or auto-regulated by itself. The next step of our study consisted of inspecting the potential controlling effect of a sigma factor to other sigma factors. In this case, only the expression profiles of sigma factors served as an input to the model (Eq. 1.).

The computed transcriptional controls between two sigma factors (listed in Table 1, the second column) are represented by arrows in Figure 3. For the measured sigma factor profiles that were highly correlated, trivial mutual interactions were found (two black arrows with opposing directions in Figure 3), which suggests that from the kinetic point of view, the regulation possibly occurred in both directions. The method used in this study did not allow us to distinguish which of them is regulator and which one is the target. The interpretation of this type of interaction is ambiguous, and further information is needed. For example, for the SigL – SigB pair, mutual regulation was proposed in this study. In the literature, SigB was suggested to be a regulator of *sigL* by Lee et al (45) using different experimental approaches. Thus, by integrating the existing information, the regulation of *sigL* by SigB during *S. coelicolor* germination was much more probable compared to the SigL-*sigB* control. Incorporation of external knowledge allowed selecting the more probable of two kinetically equivalent relations. A similar case occurred for SigR – HrdD mutual regulation, where SigR was previously suggested to regulate *hrdD* (33,35)(further discussed in paragraph 3.5.2).

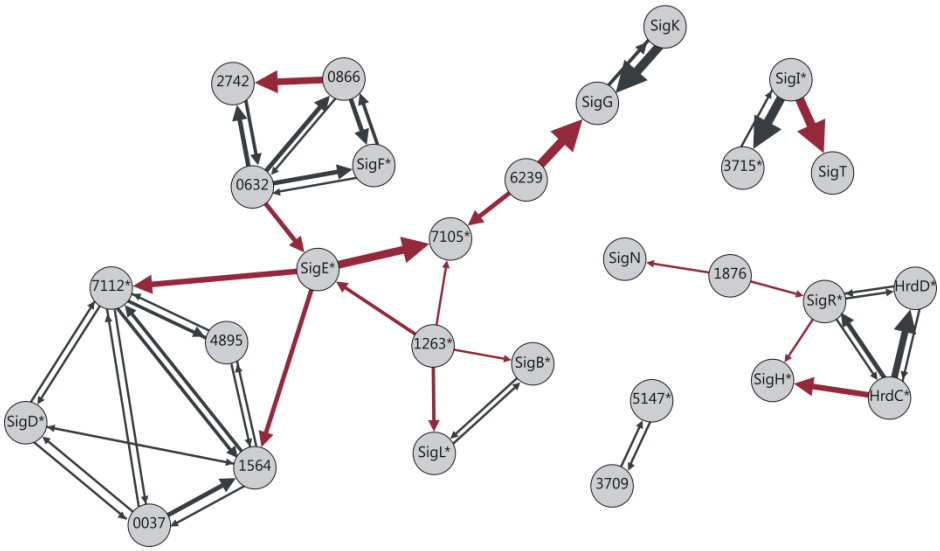


Figure 3. Possible transcriptional regulation between sigma factors (grey nodes). The thickness of the arrows is proportional to the parameter of the model w (Eq. 1) and corresponds to the regulatory effect strength. The sigma factors are indicated by name or SCO number; those with high overall expression (above 3rd quartile, category IV.) are marked with an asterisk. The black arrows correspond to mutual regulation (see text for details). The red arrows indicate individual regulation between regulators.

3.4.3 **Global Functional analysis of a time series**

Under the generally accepted assumption that co-expressed genes characterize a specific functional group, we examined gene members of kinetic clusters that were proposed in previous paragraphs for their membership in different metabolic groups. According to the database annotating the *S. coelicolor* genome (ftp://ftp.sanger.ac.uk/pub/S_coelicolor/classwise.txt), each gene was categorized into a functional or a potential functional class. The idea was to identify significantly overrepresented gene functional groups in the kinetic clusters and thus characterize individual clusters. Knowing sigma factors that control the kinetic cluster, specific cell metabolic processes were characterized as controlled by individual sigma factors.

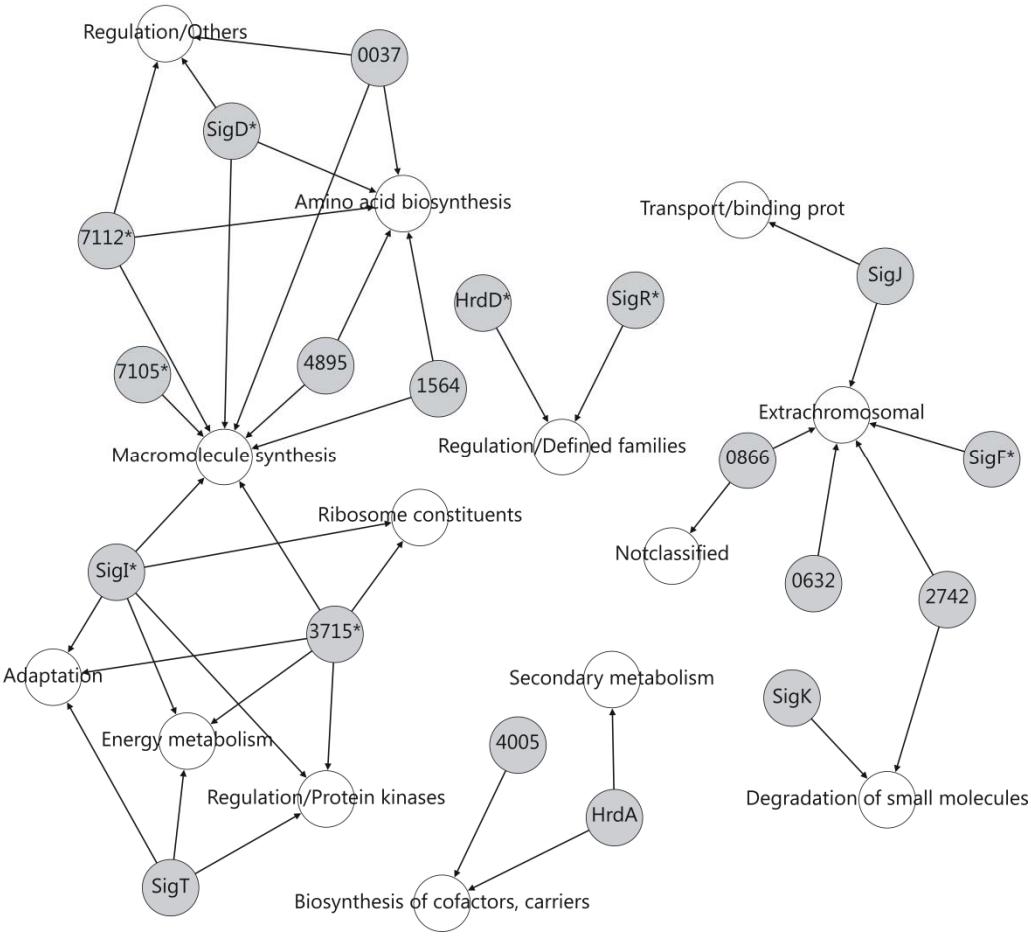


Figure 4. Suggested regulation of significantly enriched functional classes (white nodes) by sigma factors (gray nodes). The asterisk marks high overall expression of sigma factors from category IV (Figure 1A).

All highly expressed genes were assigned to kinetic clusters based on the value of the correlation coefficient between the given gene expression profile and the core profile of each cluster. Evidently, the choice of the criterion value may significantly affect the gene composition of the clusters; therefore, the choice of the criterion influences the functional characteristics of the group. Hence, for the enrichment of functional groups in the kinetic clusters, we tested four levels of the criterion (correlation coefficient > 0.7, 0.75, 0.8 and 0.85). Obviously, the number of genes assigned into the clusters differed with a different correlation criterion, but the changes in significantly overrepresented functional groups were minor (data not shown). Therefore, in contrast with the initial assumption, the value of the correlation coefficient was not crucial for the resulting functional characteristic of the cluster. As

a selection criterion, the correlation coefficient was set to be >= 0.8 (3586 gene profiles assigned into kinetic clusters) Supplementary Table 2. The coefficient can be understood as a level of membership of the gene in the particular cluster. The higher the coefficient, the more kinetically similar was the gene profile to the cluster core profile; therefore, the proposed regulation by the sigma factor had higher probability compared to genes with lower correlation coefficients.

To determine the significantly enriched gene functional groups in the kinetic clusters in comparison with the entire gene dataset, statistical Fishers' exact test was used. The functional group within the cluster was considered significantly enriched if Fishers' test p-value < 0.05, fold-change > 2 and the number of genes involved in the particular functional group in the cluster was > 7. The regulation of gene functional groups is depicted in Figure 4 (listed in Table 1, the third column). Grey nodes label sigma factors while the white nodes represent specific gene functional groups. Similar to the case in Figure 2 and Figure 3, multiple arrows pointing to the "functional nodes" represent all controlling alternatives but not simultaneous controls.

Considerable enrichment was found for the macromolecule synthesis gene functional class, which was significantly overrepresented in 4 different kinetic clusters; therefore, several sigma factors were suggested to possibly regulate the group Figure 4. Germinating spores have an urgent need for new macromolecules. It was reported (10,12) that just after germination initiation, a rapid boost of translation machinery begins. Translation is expectedly connected with ribosome constituents and amino acid biosynthesis classes of genes, which are necessary components for proteosynthesis. The other enriched gene functional groups included those associated with primary metabolism, such as genes belonging to the energy metabolism, regulations and adaptation groups.

3.5 Biological interpretation

3.5.1 Identification of the genes controlled by SigR

As suggested in previous studies (31-34), SigR controls the oxidative stress response induced by diamide, the compound that disrupts intracellular redox homeostasis by thiol oxidation. SigR directly regulates the expression of genes that help to restore redox balance and protect the cell against chemically induced

oxidations. Later studies revealed that SigR also induces target genes that participate in protein quality control, indicating that SigR regulates the cell response to protein misfolding and aggregation, also caused by the increased oxidation of enzymes induced by diamide (35). This finding is particularly important to the germination studied in this work because in dormant cells, proteins are stored as immobile aggregates (46). Further studies of the SigR regulon used the ChIP-chip assay to identify SigR binding sites under thiol-oxidative stress conditions (20), leading to suggestion of large number of SigR target genes.

From the proposed regulon of the SigR target gene (20), 145 were highly expressed in our dataset. Individual kinetic profiles of this set were tested by applying the kinetic transcriptional model (Eq.1) to determine whether, from the kinetic perspective, the suggested regulation is plausible. Verification of the ChIP-chip results by employing kinetic modeling was justified by the fact that the ChIP-chip assay provides only static information about SigR binding. Although binding may be observed, a functional relationship does not necessarily occur (7,8). Furthermore, several promoters can be bound by various sigma factors under different experimental conditions and developmental phases. Therefore, the verification of the kinetic plausibility of the interaction is essential.

The agreement between the ChIP-chip results and kinetic modeling was observed for approximately one-third of the ChIP-chip suggested target gene set. As shown in Figure 4, the highest portions from the confirmed regulations belonged to the genes whose products have the following known/predicted function: protein degradation 17% (*clpP1* - SCO2619, *clpP2* - SCO2618, *clpC* - SCO3373, *clpX* - SCO2617, *prcA* - SCO1643, *mpa* - SCO1648, and *pepN* - SCO2643), transcriptional regulators 15% (*rsrA* - SCO5217, *ndgR* - SCO5552, *rsrA2* - SCO3451, *sigR2* - SCO3450, SCO1619, and SCO7140), thiol homeostasis 13% (*rifO* - SCO7632, *trxA* - SCO3890, *trxB* - SCO3890, *trxC* - SCO0885, *mca* - SCO4967, and *trxA4* - SCO1084), oxidoreductases 11% and cofactor metabolism 11%. Regulation for genes with other functions were confirmed for only a few (under 9%). Our approach did not confirm any regulation of genes involved in energy metabolism and identified only one target gene from the lipid metabolism group and 3 from the modulation of ribosomal constituents or translation group, although these groups represent 23% of target genes identified in original ChIP-chip-based approach. The

full list of kinetically plausible SigR target genes in germination is shown in Supplementary Table 3.

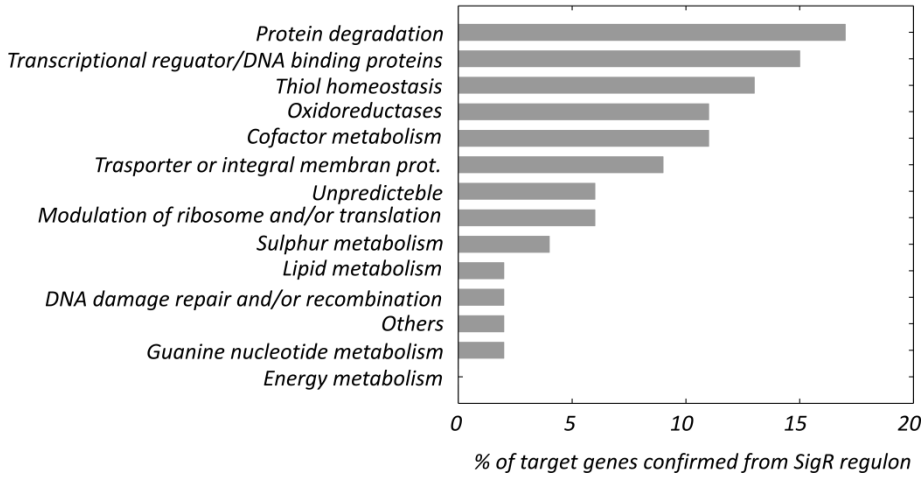


Figure 4. The distribution of SigR target genes that satisfied the kinetic transcriptional model among functional categories during germination. Functional assignments were adopted from the study of Kim et al., 2012.

Generally, kinetically confirmed SigR transcriptional controls belong to genes with “special” functions in redox homeostasis, regulation or protein quality control rather than “basal” metabolism functions, such as energy metabolism, lipid metabolism and ribosome constituents/translation. When interpreting the results of our experiments, the expression of SigR was not “arbitrarily” enhanced by adding diamide or any other chemical compounds inducing oxidative stress, unlike in all previous studies, where the expression of SigR and its regulon was induced by diamide. Additionally, this study used cells undergoing germination; however, previous studies utilized cells at later life phases. When we considered that the sigma factor activity and the selection of promoters depended on growth conditions and developmental stage, the inconsistencies in our findings and the previous studies can originate from both a difference in SigR transcriptional control under “normal” physiological growth conditions (germination) and under induced thiol-oxidative stress and from silent binding events.

We can conclude that during germination of *S. coelicolor*, the SigR regulon exhibited a similar response as observed under thiol-oxidative stress which has not been reported yet. We can speculate that a probable trigger of the stress that

consequently induced SigR expression was a stimulus provided by high content of aggregated and/or misfolded proteins present in the dormant cell from the sporulation phase. These suggestions are in agreement with previous observations by Kalifidas et. al. (35) where the protein misfolding and aggregation were caused by diamide resulting in an involvement of the SigR regulator in the response to the stress. In addition, the expression of several chaperones and protein modifiers, important for protein reactivation and quality control, were detected at protein level immediately after germination initiation (10,12). Here, for the first time, we were able to suggest a function of a sigma factor that acts during germination in *Streptomyces*.

3.5.2 **Identification of the genes controlled by HrdD**

As shown in Figure 1B, the principal sigma factor *hrdD* had the highest expression level during *S. coelicolor* germination from all sigma factors; its expression was approximately 20-fold higher than the average expression level of all detected genes. Unfortunately, the specific function of HrdD in *S. coelicolor* is unknown.

Due to the *hrdD* enormous expression, we searched for potential HrdD target genes by applying the Eq. 1 model in an one-by-one interaction manner to obtain more precise image of kinetically plausible HrdD targets than we reached on a global scale level (paragraph 3.4.3.). In total, 88 genes met the required criterion (correlation between modeled and measured profiles > 0.8). The list of these genes and their functions are shown in Supplementary Table 4. Interestingly, 30 of the 88 HrdD potential target genes were assigned a regulatory function. More than 60% of these regulatory genes belong to defined regulatory families, such as transcription regulators of the TetR, MarR, LysR, and GntR families. HrdD was also found to be a potential regulator of *hrdC* and SCO0781 coding anti-sigma factor antagonist and *hrdC*. These results are consistent with the above-mentioned investigation of global-scale functional analysis (paragraph 0) where HrdD was also proposed to control the overrepresented regulation/defined families functional group from cluster 8.

Among the predicted HrdD target genes, three genes from the DnaK-HspR regulon were found; *dnaJ* (SCO3669), *hspR* (SCO3668) coding autoregulatory repressor protein and *clpB* (SCO3661) coding ATP-dependent protease) (47). More than a third of the HrdD target genes identified here belonged to a wide group of

regulators. This finding suggests that HrdD may be a novel candidate sigma factor with a global regulatory role in *S. coelicolor* germination.

Among the identified HrdD target genes, several "trivial" regulations (see paragraph 3.4.1) were found. In the list of HrdD target genes (Supplementary Table 4), the trivial regulations are marked with a hash. For seven of these genes, the possibility of the existence of common regulators should be considered. For these genes, we previously identified possible regulation by SigR (paragraph 3.4.1), and previous binding experiments (20) also suggest that SigR is their regulator. Moreover, HrdD was suggested to be a target gene of SigR by S1 nuclease mapping and microarray experiments (33,35). In contrast, additional experiments using ChIP-chip (20) did not demonstrate the binding of SigR to the *hrdD* promoter; therefore, the interaction remains unclear. Both *hrdD* and *sigR* were among 5% of the most expressed genes during *S. coelicolor* germination and are candidates for further detailed experimental research.

3.5.3 **Comparison with suggested regulations from the literature**

Although the majority of *S. coelicolor* sigma factors have not been studied on a genomic scale in a systematic manner, many references mention individual genes that are regulated by sigma factors. Predicted sigma factor target genes and the references of these studies are arranged in the "Database of transcriptional regulation in *Streptomyces coelicolor* and its closest relatives" (DBSCR) http://dbscr.hgc.jp/index.html. We investigated the possibility of transcriptional regulation from the kinetic perspective, using the computational model (Eq.1.) for each pair of sigma factor-target gene that was suggested by DBSRC and that was among the highly expressed genes.

Altogether, 74 regulatory relationships were tested, but only 7 satisfied the kinetic model. The verified genes are shown in ***Table 2*** together with the list of the sigma factors and the number of tested genes identified in the literature. For SigB, the kinetic model fit for regulation of the small hydrophobic protein (SCO2372), *ssgC* (SCO7289), a *ssgA*-like gene encoding sporulation protein controlling septum site initiation and DNA segregation in spores (48) and two regulatory genes (*sigL* (SCO7278) and its putative anti-sigma factor SCO7277 (49)). All the candidates were identified using the consensus promoter sequence and then verified by S1

nuclease mapping (49). Our computational analysis designated genes *cwgA* (SCO6179) and *dagA* (SCO3471) as a target genes of SigE. CwgA is the first gene of the *cwg* operon involved in biosynthesis of cell wall glycans (50), and *dagA* is the gene that codes for extracellular agarase (51).

Several factors may explain why the overlap between regulations suggested in the literature and our kinetic-based approach to be relatively low. First, the regulatory set was based on various papers (more than 70) describing distinct biological phenomenon under different experimental conditions and during different life phases of *S. coelicolor*. Second, numerous studies have identified the regulation of genes involved in antibiotic production or sporulation; however, these genes are highly unlikely to play an important role in germination and thus, expectedly, this regulation may not be consistent with our dataset. Last, transcriptional control involves both the binding of various sigma factors to a single promoter and the recognition of a single promoter by various sigma factors with overlapping promoter specificities (52-54). The interaction of promoter-sigma factor depends on the developmental stage and the experimental conditions. Promoter specificity overlap can be especially useful under various stress conditions, that ultimately lead to the same type of physiological stress. The potential for promoter-sigma factor multiple responses also caused few genes from the tested set to be identified; therefore, in DBSRC a regulation by more sigma factors was proposed. It should be emphasized here that germination is a specific life period that may also require specific regulation and yet non-sudied sigma factors. Our data support this idea.

4 CONCLUSIONS

In this work a large experimental dataset containing thousands of gene temporal expression profiles that were relatively densely sampled was created. The generated dataset can serve as a source material for both further computational analyses employing time series expression data and consequent experimental studies.

We analyzed the whole-genome transcriptome dynamics during the transition phase of *S. coelicolor* from dormancy to vegetative growth. Using computational modeling, we identified the target genes and gene kinetic clusters of 29 sigma factors (out of 45 studied) and suggested potential transcriptional regulatory networks that are controlled by these sigma factors.

Specifically, we chose germination because it has not been studied sufficiently, although it influences further development of the cell. Germination also represents a system with a well-defined origin, which is suitable for numerical kinetic modeling applications. We showed, together with previous studies on computational kinetic modeling (2,5), that kinetic analysis of gene expression time series allows identifying gene control networks on a global scale. Although gene and protein expression levels may differ substantially, the dynamics at the transcript level is quite well correlated with the dynamics in protein level (3). Therefore, the time series of protein expression have been often replaced with relatively straight forwardly measured transcriptome profiles.

From the analysis of the functional groups of the target genes, we identified sigma factors that are probable regulators of basic metabolic processes activated during *S. coelicolor* germination. For a single sigma factor – SigR, the kinetic data were complemented with ChIP-chip experiment and the results were compared with the published data. We suggest particular role for the alternative principal sigma factor, HrdD, whose expression at the mRNA level was extremely high during germination in comparison with all the other genes.

We are aware that gene expression control in *Streptomyces* is more complicated than the presented model in its current form can describe. Existence of anti sigma factors or even anti-anti sigma factors document complex nature of the transcriptional control, making inference of the control networks in this organism very complicated. In principal two strategies for inference of gene expression control networks can be used. First, traditional, inspects sigma factors individually and experimentally searches for their targets. Such approach has brought most of the knowledge about *Streptomyces* gene expression control available so far. However, predominantly used gene deletion and consequent pair vice comparison of mutant and wild type strains is complicated by the inability to distinguishing between direct and transmitted control, that, in order to be resolved, require additional set of experiments, complicating their interpretation. The existence of more than sixty sigma factors in *S. coelicolor* and thousands potential target genes generate hundreds of thousands of potential regulator-target gene interactions. Their complete experimental inspection using traditional methods is virtually impossible. From such a pool only a few, were picked by other researches for detailed experimental inspections. To reconstruct the networks on a global scale from such individual

results which were often obtained independently under, sometimes, very different experimental conditions, is therefore difficult or even impossible, their quantitative features are from such data completely unfeasible. A computational modeling utilizing complex parallel kinetic data can help to overcome this hurdle, giving possibility to retain the parallel nature of the data and keep the consistency given by one experimental setup common for whole dataset. Incorporating extra information such as DNA binding data to the model, as here for the case of SigR, can contribute even more to the network inference by excluding those interactions that are physically impossible and make thus the model based predictions more accurate.

We see the contribution of this paper in providing an overview of the gene expression networks active during the studied process rather than in giving ultimate answers concerning individual interactions. Our approach reduced thousands of potential regulatory interactions to tens that can be experimentally verified and gave a global outlook on the system level of control that gives a complex picture of the whole system

Last but not least, the model presented here allows simulating kinetics of gene expression and provide possibility to make virtual experiments which can, again, point out additional experiments. Such iterative process will lead to creation of a functional model of gene expression control network that can be used to get deeper insight into the dynamic properties of gene expression control of the studied process and the topology of the network.

We are convinced that such systems level approach combining prior knowledge with computational modeling can identify, from a global perspective, regulatory networks controlling cellular processes, not only in germination and *S. coelicolor* but also in other organisms.

5 SUPPLEMENTARY DATA

Supplementary Table 1. The list of highly expressed sigma factors.

Supplementary Table 2. The list of genes assigned to the clusters.

Supplementary Table 3. Genes that represent a subset of SigR regulated genes.

Supplementary Table 4. The list of kinetically plausible HrdD target genes.

6 FUNDING

This work was supported by grants from the Czech Science Foundation [P302-11-0229 to JV] and a grant from the Grant Agency of the Charles University under contract no. [17409 to ES]. Funding for open access charge: Czech Science Foundation

7 REFERENCES

1. Bansal, M., Gatta, G.D. and di Bernardo, D. (2006) Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. *Bioinformatics*, **22**, 815-822.
2. Penfold, C.A. and Wild, D.L. (2011) How to infer gene networks from expression profiles, revisited. *Interface Focus*, **1**, 857-870.
3. de Sousa Abreu, R., Penalva, L.O., Marcotte, E.M. and Vogel, C. (2009) Global signatures of protein and mRNA expression levels. *Mol Biosyst*, **5**, 1512-1526.
4. Gao, P., Honkela, A., Rattray, M. and Lawrence, N.D. (2008) Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities. *Bioinformatics*, **24**, i70-75.
5. Honkela, A., Girardot, C., Gustafson, E.H., Liu, Y.H., Furlong, E.E., Lawrence, N.D. and Rattray, M. (2010) Model-based method for transcription factor target identification with limited data. *Proc Natl Acad Sci U S A*, **107**, 7793-7798.
6. Titsias, M.K., Honkela, A., Lawrence, N.D. and Rattray, M. Identifying targets of multiple co-regulating transcription factors from expression time-series by Bayesian model comparison. *BMC Syst Biol*, **6**, 53.
7. MacQuarrie, K.L., Fong, A.P., Morse, R.H. and Tapscott, S.J. (2011) Genome-wide transcription factor binding: beyond direct target regulation. *Trends Genet*, **27**, 141-148.
8. To, C.C. and Vohradsky, J. (2010) Measurement variation determines the gene network topology reconstructed from experimental data: a case study of the yeast cyclin network. *FASEB J*, **24**, 3468-3478.
9. Mikulik, K., Bobek, J., Bezouskova, S., Benada, O. and Kofronova, O. (2002) Expression of proteins and protein kinase activity during germination of aerial spores of Streptomyces granaticolor. *Biochem Biophys Res Commun*, **299**, 335-342.
10. Bobek, J., Halada, P., Angelis, J., Vohradsky, J. and Mikulik, K. (2004) Activation and expression of proteins during synchronous germination of aerial spores of Streptomyces granaticolor. *Proteomics*, **4**, 3864-3880.
11. Piette, A., Derouaux, A., Gerkens, P., Noens, E.E., Mazzucchelli, G., Vion, S., Koerten, H.K., Titgemeyer, F., De Pauw, E., Leprince, P. *et al.* (2005) From dormant to germinating spores of Streptomyces coelicolor A3(2): new perspectives from the crp null mutant. *J Proteome Res*, **4**, 1699-1708.
12. Strakova, E., Bobek, J., Zikova, A., Rehulka, P., Benada, O., Rehulkova, H., Kofronova, O. and Vohradsky, J. (2013) Systems Insight into the Spore Germination of Streptomyces coelicolor. *J Proteome Res*, **12**, 525-536.
13. 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.

14. Panek, J., Bobek, J., Mikulik, K., Basler, M. and Vohradsky, J. (2008) Biocomputational prediction of small non-coding RNAs in Streptomyces. *BMC Genomics*, **9**, 217.
15. Swiercz, J.P., Hindra, Bobek, J., Haiser, H.J., Di Berardo, C., Tjaden, B. and Elliot, M.A. (2008) Small non-coding RNAs in Streptomyces coelicolor. *Nucleic Acids Res*, **36**, 7240-7251.
16. Bibb, M. and Hesketh, A. (2009) Chapter 4. Analyzing the regulation of antibiotic production in streptomycetes. *Methods Enzymol*, **458**, 93-116.
17. Bucca, G., Laing, E., Mersinias, V., Allenby, N., Hurd, D., Holdstock, J., Brenner, V., Harrison, M. and Smith, C.P. (2009) Development and application of versatile high density microarrays for genome-wide analysis of Streptomyces coelicolor: characterization of the HspR regulon. *Genome Biol*, **10**, R5.
18. den Hengst, C.D., Tran, N.T., Bibb, M.J., Chandra, G., Leskiw, B.K. and Buttner, M.J. (2010) Genes essential for morphological development and antibiotic production in Streptomyces coelicolor are targets of BldD during vegetative growth. *Mol Microbiol*, **78**, 361-379.
19. Nieselt, K., Battke, F., Herbig, A., Bruheim, P., Wentzel, A., Jakobsen, O.M., Sletta, H., Alam, M.T., Merlo, M.E., Moore, J. *et al.* (2010) The dynamic architecture of the metabolic switch in Streptomyces coelicolor. *BMC Genomics*, **11**, 10.
20. Kim, M.S., Dufour, Y.S., Yoo, J.S., Cho, Y.B., Park, J.H., Nam, G.B., Kim, H.M., Lee, K.L., Donohue, T.J. and Roe, J.H. (2012) Conservation of thiol-oxidative stress responses regulated by SigR orthologues in actinomycetes. *Mol Microbiol*, **85**, 326-344.
21. Buttner, M.J. and Lewis, C.G. (1992) Construction and characterization of Streptomyces coelicolor A3(2) mutants that are multiply deficient in the nonessential hrd-encoded RNA polymerase sigma factors. *J Bacteriol*, **174**, 5165-5167.
22. Cho, Y.H., Lee, E.J., Ahn, B.E. and Roe, J.H. (2001) SigB, an RNA polymerase sigma factor required for osmoprotection and proper differentiation of Streptomyces coelicolor. *Mol Microbiol*, **42**, 205-214.
23. Sevcikova, B., Benada, O., Kofronova, O. and Kormanec, J. (2001) Stress-response sigma factor sigma(H) is essential for morphological differentiation of Streptomyces coelicolor A3(2). *Arch Microbiol*, **177**, 98-106.
24. Mao, X.M., Zhou, Z., Hou, X.P., Guan, W.J. and Li, Y.Q. (2009) Reciprocal regulation between SigK and differentiation programs in Streptomyces coelicolor. *J Bacteriol*, **191**, 6473-6481.
25. Potuckova, L., Kelemen, G.H., Findlay, K.C., Lonetto, M.A., Buttner, M.J. and Kormanec, J. (1995) A new RNA polymerase sigma factor, sigma F, is required for the late stages of morphological differentiation in Streptomyces spp. *Mol Microbiol*, **17**, 37-48.
26. Dalton, K.A., Thibessard, A., Hunter, J.I. and Kelemen, G.H. (2007) A novel compartment, the 'subapical stem' of the aerial hyphae, is the location of a sigN-dependent, developmentally distinct transcription in Streptomyces coelicolor. *Mol Microbiol*, **64**, 719-737.
27. Chater, K.F., Bruton, C.J., Plaskitt, K.A., Buttner, M.J., Mendez, C. and Helmann, J.D. (1989) The developmental fate of S. coelicolor hyphae depends upon a gene product homologous with the motility sigma factor of B. subtilis. *Cell*, **59**, 133-143.
28. Bibb, M.J., Molle, V. and Buttner, M.J. (2000) sigma(BldN), an extracytoplasmic function RNA polymerase sigma factor required for aerial mycelium formation in Streptomyces coelicolor A3(2). *J Bacteriol*, **182**, 4606-4616.

29. Mao, X.M., Zhou, Z., Cheng, L.Y., Hou, X.P., Guan, W.J. and Li, Y.Q. (2009) Involvement of SigT and RstA in the differentiation of Streptomyces coelicolor. *FEBS Lett*, **583**, 3145-3150.
30. Paget, M.S., Chamberlin, L., Atrih, A., Foster, S.J. and Buttner, M.J. (1999) Evidence that the extracytoplasmic function sigma factor sigmaE is required for normal cell wall structure in Streptomyces coelicolor A3(2). *J Bacteriol*, **181**, 204-211.
31. Paget, M.S., Kang, J.G., Roe, J.H. and Buttner, M.J. (1998) sigmaR, an RNA polymerase sigma factor that modulates expression of the thioredoxin system in response to oxidative stress in Streptomyces coelicolor A3(2). *EMBO J*, **17**, 5776-5782.
32. Kang, J.G., Paget, M.S., Seok, Y.J., Hahn, M.Y., Bae, J.B., Hahn, J.S., Kleanthous, C., Buttner, M.J. and Roe, J.H. (1999) RsrA, an anti-sigma factor regulated by redox change. *EMBO J*, **18**, 4292-4298.
33. Paget, M.S., Molle, V., Cohen, G., Aharonowitz, Y. and Buttner, M.J. (2001) Defining the disulphide stress response in Streptomyces coelicolor A3(2): identification of the sigmaR regulon. *Mol Microbiol*, **42**, 1007-1020.
34. Park, J.H. and Roe, J.H. (2008) Mycothiol regulates and is regulated by a thiol-specific antisigma factor RsrA and sigma(R) in Streptomyces coelicolor. *Mol Microbiol*, **68**, 861-870.
35. Kallifidas, D., Thomas, D., Doughty, P. and Paget, M.S. (2010) The sigmaR regulon of Streptomyces coelicolor A32 reveals a key role in protein quality control during disulphide stress. *Microbiology*, **156**, 1661-1672.
36. Lee, E.J., Cho, Y.H., Kim, H.S., Ahn, B.E. and Roe, J.H. (2004) Regulation of sigmaB by an anti- and an anti-anti-sigma factor in Streptomyces coelicolor in response to osmotic stress. *J Bacteriol*, **186**, 8490-8498.
37. Sevcikova, B., Rezuchova, B., Homerova, D. and Kormanec, J. (2010) The anti-anti-sigma factor BldG is involved in activation of the stress response sigma factor sigma(H) in Streptomyces coelicolor A3(2). *J Bacteriol*, **192**, 5674-5681.
38. Vohradsky, J. (2001) Neural network model of gene expression. *Faseb J*, **15**, 846-854.
39. Krasny, L., Tiserova, H., Jonak, J., Rejman, D. and Sanderova, H. (2008) The identity of the transcription +1 position is crucial for changes in gene expression in response to amino acid starvation in Bacillus subtilis. *Mol Microbiol*, **69**, 42-54.
40. Edgar, R., Domrachev, M. and Lash, A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, **30**, 207-210.
41. Theodoridis, S. and Koutroumbas, K. (1999) *Pattern Recognition*. Academic Press; p.557-561.
42. Vu, T.T. and Vohradsky, J. (2007) Nonlinear differential equation model for quantification of transcriptional regulation applied to microarray data of Saccharomyces cerevisiae. *Nucleic Acids Res*, **35**, 279-287.
43. Vu, T.T. and Vohradsky, J. (2009) Inference of active transcriptional networks by integration of gene expression kinetics modeling and multisource data. *Genomics*, **93**, 426-433.
44. Spall, J.C. (2003) *Introduction to Stochastic Search and Optimization*. Wiley-Interscience.
45. Lee, E.J., Karoonuthaisiri, N., Kim, H.S., Park, J.H., Cha, C.J., Kao, C.M. and Roe, J.H. (2005) A master regulator sigmaB governs osmotic and oxidative response as well as differentiation via a network of sigma factors in Streptomyces coelicolor. *Mol Microbiol*, **57**, 1252-1264.

46. Cowan, A.E., Koppel, D.E., Setlow, B. and Setlow, P. (2003) A soluble protein is immobile in dormant spores of Bacillus subtilis but is mobile in germinated spores: implications for spore dormancy. *Proc Natl Acad Sci U S A*, **100**, 4209-4214.
47. Bucca, G., Brassington, A.M., Hotchkiss, G., Mersinias, V. and Smith, C.P. (2003) Negative feedback regulation of dnaK, clpB and lon expression by the DnaK chaperone machine in Streptomyces coelicolor, identified by transcriptome and in vivo DnaK-depletion analysis. *Mol Microbiol*, **50**, 153-166.
48. Noens, E.E., Mersinias, V., Traag, B.A., Smith, C.P., Koerten, H.K. and van Wezel, G.P. (2005) SsgA-like proteins determine the fate of peptidoglycan during sporulation of Streptomyces coelicolor. *Mol Microbiol*, **58**, 929-944.
49. Lee, E.J., Cho, Y.H., Kim, H.S. and Roe, J.H. (2004) Identification of sigmaB-dependent promoters using consensus-directed search of Streptomyces coelicolor genome. *J Microbiol*, **42**, 147-151.
50. Hong, H.J., Paget, M.S. and Buttner, M.J. (2002) A signal transduction system in Streptomyces coelicolor that activates the expression of a putative cell wall glycan operon in response to vancomycin and other cell wall-specific antibiotics. *Mol Microbiol*, **44**, 1199-1211.
51. Buttner, M.J., Smith, A.M. and Bibb, M.J. (1988) At least three different RNA polymerase holoenzymes direct transcription of the agarase gene (dagA) of Streptomyces coelicolor A3(2). *Cell*, **52**, 599-607.
52. Viollier, P.H., Kelemen, G.H., Dale, G.E., Nguyen, K.T., Buttner, M.J. and Thompson, C.J. (2003) Specialized osmotic stress response systems involve multiple SigB-like sigma factors in Streptomyces coelicolor. *Mol Microbiol*, **47**, 699-714.
53. Roth, V., Aigle, B., Bunet, R., Wenner, T., Fourrier, C., Decaris, B. and Leblond, P. (2004) Differential and cross-transcriptional control of duplicated genes encoding alternative sigma factors in Streptomyces ambofaciens. *J Bacteriol*, **186**, 5355-5365.
54. Paget, M.S.B., Hong, H.J., Bibb, M.J. and Buttner, M.J. (2002) In Hodgson, D. A. and Thomas, C. M. (eds.), *Signals, Switches, Regulons, and Cascades: Control of Bacterial Gene Expression* Cambridge University Press.

8 TABLES

Table 1. The table summarizes results obtained by global kinetic analysis of microarray time series (paragraph 3.4) from *S. coelicolor* germination. For tested sigma factors potential regulations were suggested for: the group of genes – 2nd column, the target sigma factor genes – 3rd column and for the functional classes (as annotated in Sanger database) – 4th column. The individual members of kinetic clusters are listed in the Supplementary Table 2. Visualization of the 2nd column is in Figure 2, the 3rd column in Figure 3 and the 4th column in Figure 4.

Table 1.

Regulator (sigma facotor)	Suggested regulation of kinetic cluster	Suggested target sigma factor gene	Suggested regulated gene functional class
HrdA	CL 11 CL 14 CL 37		Biosynthesis of cofactors, carriers Secondary metabolism
HrdC	CL 3 CL 5	*hrdD* *sigH* *sigR*	
HrdD	CL 8	*hrdC* *sigR*	Regulation/Defined families
SigB	CL 30	*sigL*	
SigD	CL 36	SCO0037 SCO1564 SCO7112	Amino acid biosynthesis Macromolecule synthesis Regulation/Others
SigE	CL 28	SCO1564 SCO7105 SCO7112	
SigF	CL 6 CL 22 CL 31	SCO0632 SCO0866	Extrachromosomal
SigG	CL 3	*sigK*	
SigH	CL 5		
SigI	CL 1 CL 7 CL 11	SCO3715 *sigT*	Adaptation Energy metabolism Macromolecule synthesis Regulation/Protein kinases Ribosome constituents
SigJ	CL 25		Extrachromosomal Transport/binding prot
SigK	CL 15	*sigG*	Degradation of small molecules
SigR	CL 8	*hrdC* *hrdD* *sigH*	Regulation/Defined families
SigT	CL 1 CL 13		Adaptation Energy metabolism Regulation/Protein kinases
SCO0037	CL 28 CL 36	SCO1564 SCO7112 *sigD*	Amino acid biosynthesis Macromolecule synthesis Regulation/Others
SCO0632	CL 6 CL 22	SCO0866 SCO2742 *sigE* *sigF*	Extrachromosomal

Table 1. cont.

Regulator (sigma facotor)	Suggested regulation of kinetic cluster	Suggested target sigma factor gene	Suggested regulated gene functional class
SCO0866	CL 31 CL 2 CL 16 CL 22	SCO0632 SCO2742 *sigF*	Extrachromosomal Notclassified
SCO1263	CL 18 CL 30	SCO7105 *sigB* *sigE* *sigL*	
SCO1564	CL 26 CL 28	SCO0037 SCO4895 SCO7112 *sigD*	Amino acid biosynthesis Macromolecule synthesis
SCO1876		*sigN* *sigR*	
SCO2742	CL 22 CL 32 CL 42	SCO0632	Degradation of small molecules Extrachromosomal
SCO3709		SCO5147	
SCO3715	CL 1 CL 7 CL 11	*sigI*	Adaptation Energy metabolism Macromolecule synthesis Regulation/Protein kinases Ribosome constituents
SCO4005	CL 14		Biosynthesis of cofactors, carriers
SCO4895	CL 26	SCO1564 SCO7112	Amino acid biosynthesis Macromolecule synthesis
SCO4908	CL 27		
SCO4996	CL 41		
SCO5147		SCO3709	
SCO6239	CL 3	SCO7105 *sigG*	
SCO7105	CL 9 CL 18		Macromolecule synthesis
SCO7112	CL 28 CL 36	SCO0037 SCO1564 SCO4895 *sigD*	Amino acid biosynthesis Macromolecule synthesis Regulation/Others

Table 2. The list of sigma factors and the number of their tested target genes based on references in the literature. The references can be found at the DBSCR database webpage http://dbscr.hgc.jp/index.html.

Table 2.

Sigma factor	**Gene identifier**	**Number of suggested target gene from the literature**	**Confirmed regulation for gene**	**Gene SCO nr.**
HrdB	SCO5820	34	X	X
SigB	SCO0600	16	SCO7289	7289
			SCO7277	7277
			SCO2372	2372
			sigL	7278
SigE	SCO3356	9	*dagA*	3471
			cwgA	6179
SigH	SCO5243	8	X	
HrdD	SCO3202	3	X	
LitS	SCO0194	1	X	
SigN	SCO4034	1	X	
SigF	SCO4035	1	X	
SigG	SCO7341	1	X	

Appendix B

Paper II

Systems Insight into the Spore Germination of *Streptomyces coelicolor*[2)]

Strakova, E., Bobek, J., Zikova, A., Rehulka, P., Benada, O., Rehulkova, H., Kofronova, O. and Vohradsky, J.

Journal of Proteome Research, Jan 4 2013, 12(1), 525-536.

Journal of proteome research

Article

pubs.acs.org/jpr

Systems Insight into the Spore Germination of *Streptomyces coelicolor*

Eva Strakova,[†] Jan Bobek,[†,‡] Alice Zikova,[†] Pavel Rehulka,[§] Oldrich Benada,[∥] Helena Rehulkova,[§] Olga Kofronova,[∥] and Jiri Vohradsky*[,†]

[†]Institute of Microbiology, Academy of Sciences of the Czech Republic, Laboratory of Bioinformatics, Vídeňská 1083, 142 20 Prague 4, Czech Republic

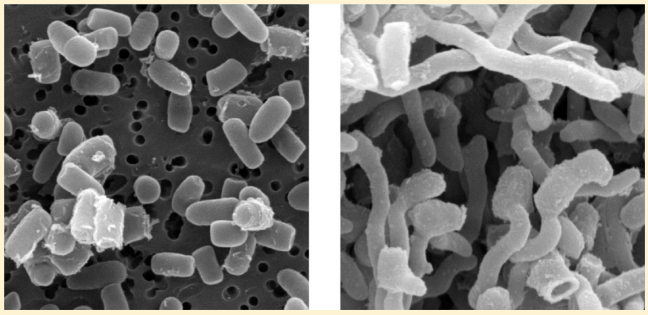
[‡]Institute of Immunology and Microbiology, First Faculty of Medicine, Charles University in Prague, Studničkova 7, 128 00 Praha 2, Czech Republic

[§]Institute of Molecular Pathology, Faculty of Military Health Sciences, University of Defence, Třebešská 1575, CZ-500 01 Hradec Králové, Czech Republic

[∥]Institute of Microbiology, Academy of Sciences of the Czech Republic, Laboratory of Molecular Structure Characterization, Vídeňská 1083, 142 20 Prague 4, Czech Republic

Supporting Information

ABSTRACT: An example of bacterium, which undergoes a complex development, is the genus of *Streptomyces* whose importance lies in their wide capacity to produce secondary metabolites, including antibiotics. In this work, a proteomic approach was applied to the systems study of germination as a transition from dormancy to the metabolically active stage. The protein expression levels were examined throughout the germination time course, the kinetics of the accumulated and newly synthesized proteins were clustered, and proteins detected in each group were identified. Altogether, 104 2DE gel images at 13 time points, from dormant state until 5.5 h of growth, were analyzed. The mass spectrometry identified proteins were separated into functional groups and their potential roles during germination were further assessed. The results showed that the full competence of spores to effectively undergo active metabolism is derived from the sporulation step, which facilitates the rapid initiation of global protein expression during the first 10 min of cultivation. Within the first hour, the majority of proteins were synthesized. From this stage, the full capability of regulatory mechanisms to respond to environmental cues is presumed. The obtained results might also provide a data source for further investigations of the process of germination.



KEYWORDS: *germination, differentiation, protein expression, Streptomyces*

INTRODUCTION

The complex life cycle of *Streptomyces coelicolor*, a Gram-positive mycelial bacterium, involves a cascade of morphologically distinguishable developmental phases with different metabolism and physiology. Filamentous growth begins with the germination of a single spore and continues with the generation of branched mycelium of vegetative hyphae and the subsequent transition from vegetative growth to form aerial hyphae, which generate unigenomic spores in long chains.[1]

Dormant spores are resistant to unfavorable conditions due to the formation of a complex coat structure that protects the cellular content and enables cell survival through dispersion. Under favorable environmental conditions, dormancy is broken and germination is induced. Various stimuli have been tested for the induction of germination, such as heat shock, mechanical disruption, and the presence of nutrient germinants,[1] however, signaling or regulatory mechanisms essential for triggering of germination remain unknown. A comprehensive insight into the mechanisms controlling *Streptomyces* germination is of primary importance both for practical applications, as *Streptomyces* species are important producers of antibiotics and other secondary metabolites, and for fundamental studies of cell development.

Recently, several studies have focused on the systems investigation of the *Streptomyces* transition between developmental phases.[2−6] However, *Streptomyces* germination has not been investigated in a systems manner yet. A proteomic approach was employed to characterize the activation of proteins and ribosomes in the transition from dormant to germinating spores in *Streptomyces granaticolor*.[7] Bobek et al. showed that aerial spores contain a major group of chaperones that modify their affinity to target proteins through reversible phosphorylation. Phosphorylation, together with the increased

Received: October 18, 2012
Published: November 26, 2012

hydration of the cytoplasm, supports the association of chaperones with proteins and ribosomes. Thus, molecular chaperones can assist in protein folding, protect nascent proteins from aggregation and activate inactive ribosomes, which are present in the spores, to their functional forms. Some *de novo* expressed proteins, such as enzymes involved in glycolysis, the citric acid cycle, amino acid and nucleic acid metabolism, and proteins of the translational apparatus, have been identified. A previous study of *S. granaticolor* in experiments using the transcription inhibitor rifamycin revealed that dormant spores preserve pre-existing mRNAs, which are expressed at the beginning of germination.[8]

It has been shown that the signaling molecule, cyclic AMP (cAMP), is important for efficient germination.[9] The effects of mutations in cAMP-binding proteins on the induction of germination in *S. coelicolor* have been studied,[9−11] and these mutants exhibited a reduction in the germination rate. Mutagenesis in cAMP receptor protein (Crp) suggested its role as a central regulatory protein for all developmental steps in the cell.[12] The *crp* mutant exhibits a decreased level of peptidoglycan hydrolase (SCO5466), which is considered to play a key role during the degradation of the spore wall in the germinating cell. Hydrolases facilitate changes in spore morphology and optical features. Both of these modifications occur during germination. The investigation of the hydrolytic enzymes involved in cell wall degradation revealed a crucial role for these enzymes in the growth and development of *S. coelicolor*.[13]

To further elucidate the process of germination in more systematic way, we employed a proteomic-based approach in a time-dependent manner. We observed dynamic changes in the proteome at 13 time points in 5.5 h of germination in *S. coelicolor* using 2D gel electrophoresis to examine both the protein concentration level (fluorescent staining) and newly expressed proteins (*in vivo* radioisotope labeling) at each time point. Proteins with increased expression were identified using mass spectrometry (MS). The results of an analysis of the protein kinetic profiles and published information concerning protein functions assessed the biochemical processes associated with initiation and other phases of germination. This proteomic study provides systems insight into the mechanism underlying *S. coelicolor* germination.

MATERIALS AND METHODS

Spore Cultivation

S. coelicolor A3(2) M145 spores were pregerminated in 2× YT media for 24 h (160 rpm, 30 °C).[1] Three milliliters of inoculum was transferred to solid agar plates (0.4% yeast extract, 1% malt extract, 0.4% glucose, 2.5% bacterial agar, pH 7.2) overlaid with cellophane discs and cultivated for 14 days at 28 °C. The harvested spores were used for germination.

Germination and Radioactive Protein Labeling

The spores were manipulated on ice prior to the induction of germination. Homogenization and mechanical disruption of the outermost coat of the spore was performed[7] using an Ultra-Turrax mixer. The suspension was subsequently washed in 10 mL of distilled water and centrifuged for 10 min (4 °C). The spore pellet was diluted in AM liquid medium (containing 18 amino acids at 0.2 mM, devoid of cysteine and methionine, containing 20 mM KH_2PO_4, 30 mM Na_2HPO_4, 2% glucose, 0.05% $MgCl_2$, 0.5 mM $CaCl_2$, 7 mM KCl, and a 0.01% mixture of bases) at a ratio of 2 mL of AM media to 0.3 g of spores. A 2-mL aliquot of the spore suspension in AM medium was used for the inoculation of 80 mL of AM media in a 500-mL flask. For boosting synchrony within the population, the spores were subjected to a 10-min heat shock treatment at 50 °C. Further germination continued at 37 °C, with shaking at 160 rpm.[1]

Newly expressed proteins were labeled *in vivo* with 30 μCi of ^{35}S Cysteine-methionine (Tran^{35}S-Label, MB Biomedicals) in 30-min radioactive pulses, except for the first time point (T0), which was labeled for 10 min during heat shock. The incubation was terminated with the addition of 2 mL of trichloroacetic acid to a final concentration of 2.5% and centrifuged for 10 min (4 °C). The resulting pellet was washed with 1% nonradioactive cysteine−methionine to remove radioactive ^{35}S Cysteine−methionine remains and stored at −20 °C.

Separate cultivations were processed for each biological replicates at individual time point. Number of replicates are given in Table 1.

Table 1. Number of 2DE Gel Images Corresponding to Biological Replicates Analyzed for Individual Time Points of the Experiment

time [h]	number of SYPRO Ruby-stained images	number of radiolabeled images
Dormant	4	0
0	4	4
0.5	4	4
1	4	4
1.5	3	3
2	4	4
2.5	3	3
3	5	5
3.5	5	5
4	5	5
4.5	5	5
5	4	4
5.5	4	4

Protein Isolation and Purification for 2D Electrophoresis

The spores were resuspended and washed in 10 mL of standard Tris buffer (pH 7.6). Each sample contained 800 μL of lyses buffer and 1 μL of benzoase from the Bacterial Protein Prep Kit (Qiagen), 1 μL of protease inhibitor (Roche, 100× diluted) and 3 μL of PMSF (phenylmethanesulfonylfluoride), and the solutions mixed and incubated on ice for 1 h. The cells were maintained in tubes containing zirconium sand and two 4-mm glass beads were used to disrupt the samples at 9 × 45 s in a FastPrep-24 machine (Biomedicals) and rechilled between each round. Finally, the samples were incubated on ice for 20 min. After centrifugation at 14 000*g* for 20 min, the supernatant was transferred to 50-mL tubes and 6 mL of acetone was added for overnight storage at −20 °C. The proteins were cleaned using the 2D Clean-Up Kit (Amersham) and dissolved in 500 μL of sample buffer (4% CHAPS, 0.8% Pharmalyte 3−10, 65 mM dithiothreitol (DTT), 8 M urea, bromphenol blue). The protein concentration was determined using the 2D Quant Kit (Amersham) in triplicates.

2D Electrophoresis and Gel Processing

A 200−290 μg sample of proteins was loaded onto 24-cm strips (Bio-Rad), pH 4−7 and focused on an isoelectric base using an IPG Phor II (Amersham). The second dimension was run on

12.5% polyacrylamide gels, 25.5 × 20.5 cm, for 4.5 h, 120 W, 20 °C, EtanDalt II (Amersham).

The gels were stained overnight with SYPRO Ruby fluorescent dye and washed 3 times (10% methanol, 7% acetic acid). The intensity of the SYPRO Ruby spots was scanned using a Phosphoimager FX (Bio-Rad). The gels were dried, and after 4 days of exposition to BAS cassettes (Fujifilm), the gels were scanned using a Phosphoimager FX to determine the protein radioactivity. The radioactive signal was proportional to the protein synthesis during the 30 min ^{35}S cysteine–methionine radioactive pulse (10 min radioactive pulse for the first time point T0). Both stained and radioactive gel images were processed and compared using software PDQuest 8.0.1 (Bio-Rad) to detect changes in the intensities for particular gel spots (proteins) in time and across replicates. Altogether, 54 2DE gels for SYPRO Ruby staining and 50 radiolabeled 2DE gels were analyzed. SYPRO Ruby stained gels and radiolabeled gels were arranged into individual matchsets. The SYPRO Ruby matchset reference gel contained 671 individual protein spots, and the radiolabeled matchset reference gel contained 404 spots. All gels were assembled into a single high level matchset whose reference gel contained a total of 782 protein spots. All visible spots were picked from a preparative gel and analyzed by mass spectrometry (see section Mass Spectrometry Analysis).

Proteomic Data Normalization

The 2D electrophoretic spot intensities in individual gels were standardized by dividing the spot intensities by the total protein concentration loaded on the gel. The multiplicative factor was calculated from accumulative gels (SYPRO Ruby). We assumed that the logarithm of intensities on accumulative gels is normally distributed.[14] The means of these distributions was distributed around a common mean. Therefore, the means of all spot distributions for all stained gels were averaged and a multiplicative factor, that adjusted all distributions to the same mean, was computed for each gel. As a single gel contains both fluorescent and radiolabeling, the multiplicative factors derived for the SYPRO Ruby-stained gels were also used for normalizing the radioactive-based images. During germination, protein synthesis linearly increases in time; distributions means for the radioactive gels were distributed along a straight line (data not shown). Therefore, the correction coefficients for the radioactive gel intensities were derived that adjusted the distribution means to this straight line. Normalized data for each spot and both types of labeling are listed in Supporting Information Table 1; for a reference gel see Supplemetary Information Figure 1.

To assess the degree of variance in quantification of relative protein abundance levels, we calculated the coefficient of variation (CV) for replicates within the individual time point for both types of labeling. The values were computed from the normalized data which were used for all further computations. Mean CV for SYPRO Ruby stained gels was 0.39, and for radiolabeled gels, it was 0.54. These values were distorted by a time dependence; the highest CV was observed for earliest gels and decreased over time (CVmax = 0.45 and CVmin = 0.27 for stained gels, and CVmax = 0.62 and CVmin = 0.4 for radiolabeled gels). These values were comparable with those previously reported in literature for 2D gel electrophoresis experiments (20–40%),[15] higher for radiolabeled gels. Higher CV for radiolabeled gels was caused mainly by the gels of first time points where proteosynthesis starts and the degree of variation is high both for technical reasons, when only small number of spots appeared on a gel, and for inherent variation among biological replicates, which is known to be high in *Streptomyces* in general. We tried to overcome this problem by increased number of gels and sample replicates which are almost 2-fold higher than in a usual proteomic experiment of this scale (see Table 1).

Relative Protein Concentration Profiles

The time series of relative protein concentrations for both protein accumulation (SYPRO Ruby) and protein synthesis (radiolabeling) were obtained by averaging the normalized intensities across replicates at specific time points. Averaging was performed only when the spot intensity was detected on more than one gel among replicates (3–5 replicates gels per time point). Before averaging, the outlier intensities among the replicates were filtered using the Q-test. The temporal profiles from the resulting protein concentrations were assessed for missing values in between two non-zero intensity levels. The sudden total disappearance of a protein is biologically implausible and could represent an experimental artifact. Therefore, the missing values were linearly interpolated.

Profiles Clustering

The kinetic profiles of protein accumulation (SYPRO Ruby staining) and protein synthesis (radiolabeling) were sorted separately into groups according to similar expression kinetic profiles. The groups with typical kinetic profiles (designated core clusters) were determined using the k-means clustering method and further profile classification was based on the correlation between the protein profiles and average core cluster profiles.

As the k-means method randomly selects a cluster seed at the beginning of the clustering procedure, the algorithm was repeated 1000 times for a predefined number of clusters (n). The groups with maximal intersect among repeated runs were identified, and only those profiles that appeared in one cluster in at least 75% of the runs were selected to form a core cluster. The 75% limit was chosen arbitrarily after visual inspection of resulting clusters.

The optimal number of clusters (n) for k-means method must be estimated; therefore, the clustering procedure was repeated for various numbers of clusters (n). For each resulting core cluster, the within-cluster sums of point-to-centroid distances were calculated and plotted as a function of the numbers of clusters (n) (data not shown). The points where the curve significantly flattens are potentially optimal numbers of clusters (n) for the given data set. The most relevant value for the protein accumulation was found to be $n = 7$, and for radiolabeling, the relevant value was $n = 6$.

To form final clusters, the profiles that were not assigned to any of the core clusters were classified into one cluster according to their correlation (Pearson) with the average core cluster profile. Only those proteins, whose correlation coefficients were higher than 0.75 were assigned, and those with coefficients <0.75 were not assigned to any of the clusters, forming a separate group. This group contained 56 SYPRO Ruby profiles (22%) and 20 radiolabeled profiles (13%) among the total number of identified protein spot expression profiles.

Singular Value Decomposition (SVD)

The protein expression profiles were arranged in a matrix $\mathbf{A}_{i\times j}$ with i rows representing individual protein spots profiles and j columns corresponding to measured time point intensities

Matrix **A** can be decomposed to the following form:

dx.doi.org/10.1021/pr300980v | J. Proteome Res. 2013, 12, 525–536

$$\mathbf{A} = \mathbf{USV}^{\mathrm{T}} \quad (1)$$

where $U_{i\times i}$ is unitary matrix, $S_{i\times j}$ is diagonal matrix with singular values λ of the matrix **A** on the diagonal and $\mathbf{V}^{\mathrm{T}}_{j\times j}$ is transpose unitary matrix with eigenvectors as rows. Each of the eigenvectors of the matrix **V** has a unique profile and can be associated with particular temporal biological process.[4] The linear combination of matrixes **U** and **V** facilitate the full reconstruction of the original matrix **A** using only a subset of eigenvectors. If required, the influence of a particular process associated with given eigenvector can be filtered from the protein expression profile. Filtering is realized by substituting 0 for a corresponding singular value in the position on the diagonal of the matrix **S** and reconstructing the data matrix **A** using eq 1. Often, the first eigenvector is associated with the growth of the population and therefore it obscures our examination of the developmental changes in the expression of individual protein spots. Therefore, the first eigenvector for the radiolabeled gels was filtered out to identify changes that occur during germination. We filtered out the first eigenvector for radiolabeled gels by rearranging the profile data matrix **A** to $\mathbf{A}_c$ according to eq 2:

$$\mathbf{A}_c = \mathbf{US}_c\mathbf{V}^T \quad (2)$$

where the matrix $\mathbf{S}_c$ was created by replacing a singular value λ_1 in matrix **S** with 0. The protein profiles (rows) in matrix $\mathbf{A}_c$ were then clustered as described above.

Protein Preparation for Mass Spectrometry

The stained protein spots were processed according to the standard protocol for mass spectrometry protein identification[16] with minor modifications. Briefly, the cut gel pieces containing separated proteins were washed with 100 μL of 100 mM NH_4HCO_3 and 400 μL of acetonitrile (5 min). The washing solution was removed and the proteins were in-gel reduced using 50 μL of 10 mM DTT in 100 mM NH_4HCO_3 (30 min at 56 °C). After the addition of 400 μL of acetonitrile and brief vortexing, the supernatant was removed. The proteins were alkylated in-gel using 50 μL of 50 mM iodoacetamide in 100 mM NH_4HCO_3 (30 min) in the dark. To terminate alkylation, the reaction solution was removed and the gel pieces were washed with 400 μL of 100 mM NH_4HCO_3 (5 min) followed by the addition of 400 μL of acetonitrile (5 min). The shrunken gel pieces were rehydrated at 4 °C (2 h) using affinity purified[17] porcine trypsin (Promega, modified, sequencing grade; about 10 ng/μL) in 10 mM NH_4HCO_3. A sufficient volume of 10 mM NH_4HCO_3 was added to completely cover the gel pieces, and the digestion was conducted overnight at 37 °C. Approximately 20 μL of 1% trifluoroacetic acid was added to extract the peptides, and the supernatant was further processed for MALDI-TOF/TOF mass spectrometry analysis.

Mass Spectrometry Analysis

The MALDI-TOF/TOF mass spectrometry measurements were performed using a 4800 Proteomics Analyzer (Applied Biosystems, Framingham, MA). The MS and MS/MS data were acquired and processed using a 4000 Series Explorer v.3.6 (Applied Biosystems). Up to 10 precursors from the MS spectra with an S/N ratio of greater than 100 were selected from a particular sample spot analysis for the MS/MS fragmentation analysis and acquisition, and sorted according to the decreasing S/N value; the contaminant peaks (keratins, trypsin autolysis, etc.) were automatically excluded from the MS/MS analysis within the interpretation method of the 4000 Series Explorer software. The isolation parameter for the precursor selection was set at 200 for the resolution of the ion-gating mechanism. The stainless steel target with 384 sample spots (with additional 13 calibration spots) and a MALDI matrix α-cyano-4-hydroxycinnamic acid (5 mg/mL) in 60% acetonitrile/0.1% TFA (v/v) were used in all MALDI experiments. The digests were purified either using stop-and-go extraction tips[18] with the subsequent addition of MALDI matrix to the sample spot containing the eluted peptides or using a matrix-tip with the direct elution of peptides and MALDI matrix on the MALDI target plate.[19] The accelerating voltage in the ion source for the MS mode was 20 kV. In the MS/MS mode, the accelerating voltage was 8 kV, which was modified after ion selection to 1 kV within the collision cell; after the ions passed the collision cell, the voltage was raised to 15 kV. Delayed extraction was performed in all experiments and optimized for m/z 2100 in the MS mode. This MALDI-TOF/TOF instrument is equipped with an Nd:YAG laser at 355 nm, with a 3−7 ns pulse and a 200-Hz firing rate. The maximum pulse energy was 20 μJ, and it was attenuated appropriately for the analysis of the samples. Both MS and MS/MS analyses in the positive mode were obtained using reflectron. The dual microchannel plate detector was set for 1.94 kV in the MS mode and 2.16 kV in the MS/MS mode. The peaks were detected using the internal algorithm of the 4000 series software with an S/N parameter set to 10 in the MS mode and 5 in the MS/MS mode, using the cluster area optimization feature.

Protein Identification

The peak lists in the Mascot generic format were generated from mass spectra using the Peaks-to-Mascot function incorporated in the 4000 Series Explorer software. The peaks from MS analysis were detected in an m/z range of 700−5000 with an S/N ratio greater than 18, whereas the detection of the MS/MS peaks with an S/N ratio greater than 9 was in the range from 68 up to an m/z value of 50 m/z units lower than the precursor m/z value. These peak lists containing both MS information from the MS run and the fragmentation data of selected precursors from the MS/MS run were submitted through Mascot Daemon (ver. 2.1.0) to the Mascot database search engine (local installation, ver. 2.1.04). The following parameters were used for the combined search (MS and MS/MS data): database, NCBInr (ver. Nov 27, 2011); taxonomy, all entries (number of sequences: 12 603 350); enzyme, trypsin; allowed missed cleavages, 1; fixed modifications, carbamidomethyl (C); variable modifications, oxidation (M), pyro-carbamidomethyl (N-term C), pyro-Glu (N-term E), pyro-Glu (N-term Q); peptide tolerance, 50 ppm; MS/MS tolerance, 300 mmu; peptide charge, (+1); monoisotopic masses; instrument, MALDI-TOF-PSD. Hits obtained with a probability lower than 0.05 to be a randomly occurring match and at least one successful peptide fragmentation confirming the identity of the protein were considered as successful protein identifications. MS identification data are stored in Supporting Information Table 2.

Scanning Electron Microscopy

Samples of dormant or germinated spores were fixed with 3% (w/v) glutaraldehyde in cacodylate buffer, pH 7.2, washed in the same buffer and allowed to settle for 48 h at 4 °C onto 0.2-μm poly-L-lysine-treated[20] SPI-pore filters (SPI supplies, West Chester, PA). The samples were dehydrated in an alcohol series (25%, 50%, 75%, 80%, 90%, 96% and 100%) followed by

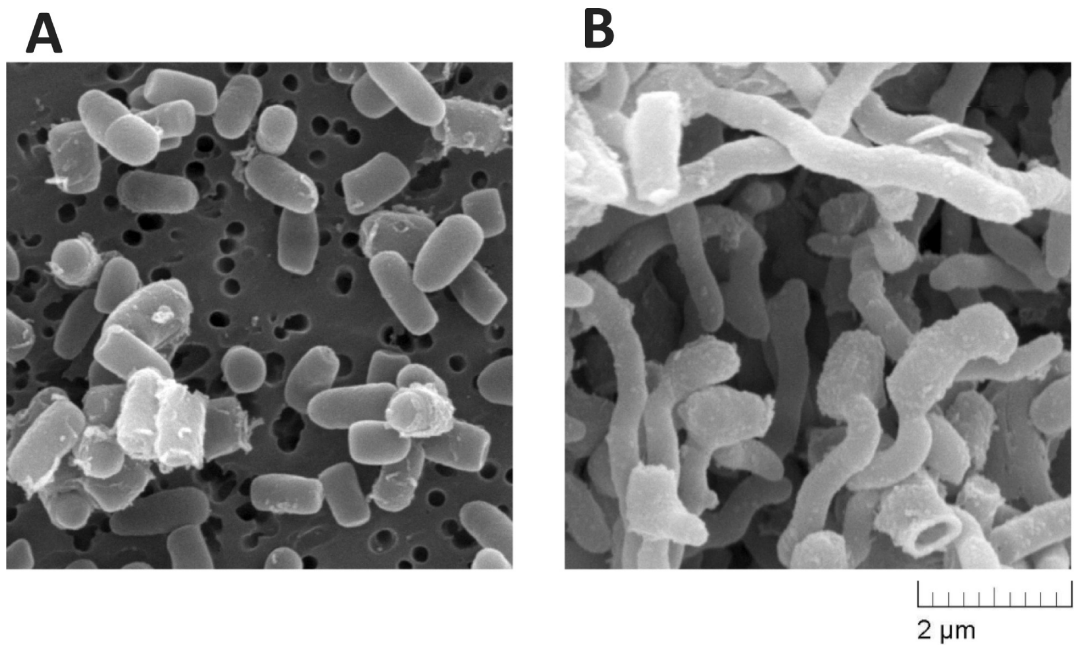


Figure 1. Spores of *S. coelicolor* observed using electron microscopy. (A) Dormant spores, − T Dorm. (B) Germinating spores at T5.5 with visible germ tubes. The images were acquired at a primary magnification of 30 000 times.

absolute acetone and dried in a critical-point device (Balzers 010, Balzers Union Ltd., Lichtenstein). The dried samples were mounted onto aluminum stubs, sputter-coated with gold (Polaron SC 510, Watford, U.K.) and examined on an Aquasem (Tescan, Brno, Czech Republic) scanning electron microscope at 15 kV in high vacuum mode. The selected samples were re-examined at a higher resolution on a Tescan Vega LSU (Tescan, Brno, Czech Republic) scanning electron microscope at 20 kV.

RESULTS AND DISCUSSION

Experimental Design

To obtain insight into the biochemistry of *S. coelicolor* germination, we examined changes in the protein accumulation and synthesis associated with this initial process of cell development. Protein synthesis was indicated by the detection of signal from radiolabeling, which specified *de novo* synthesized proteins during a radioactive pulse, whereas protein accumulation was proportional to the fluorescence signal of SYPRO Ruby, which stains the total amount of proteins in the 2D gels.

The total proteins for the 13 investigated time points were isolated from germinating spores in 3−5 repeats. As an initial point of the time courses, we used proteins from dormant spores to retain a defined start (T Dorm). The next sample (T0) was obtained immediately after the cells were subjected to a 10-min activation with 50 °C heat shock,[1] during which time the spores were incubated in presence of ^{35}S cysteine−methionine. Subsequently, 11 temporally spaced samples were collected at 30 min time intervals from 0.5 to 5.5 h (referred as T0.5−T5.5), with 30-min radiolabeling.

During the monitored 5.5 h period, the spores experienced relatively synchronous germination, which was assessed microscopically. Throughout the process, dormant spores (Figure 1A) lost hydrophobicity, became swollen, and subsequently, elongated germ tubes appeared (Figure 1B).

Time Outline of *de Novo* Synthesized Proteins

Generally, once the protein synthesis begins (recognized radioactive spot on 2D gel), the expression gradually increases with tracked time. The number of newly emerged proteins at specific time periods is depicted in Figure 2. When a protein was identified in more than one spot on the 2D gels, we further analyzed the earliest observed corresponding radioactive spot.

Figure 2 shows that the synthesis of 25 proteins is immediately invoked during heat shock (T0) and not many

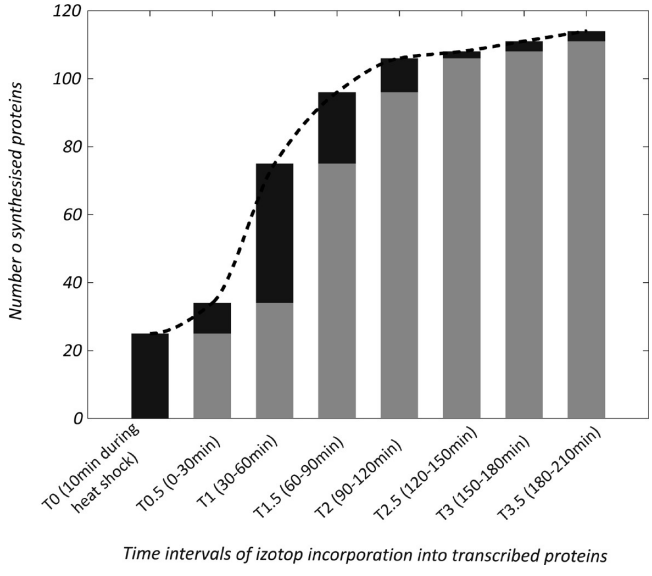


Figure 2. The columns show the numbers of synthesized (radioactive) proteins in measured time periods. The black part of the columns indicate the number of radioactive proteins whose expression is detected for the first time. Time intervals of isotope incorporation into translated or synthesized protein were 30 min; except at T0, where the incubation with radioactive label lasted 10 min during heat shock. Once initiated, the expression continued to increase with time.

dx.doi.org/10.1021/pr300980v | J. Proteome Res. 2013, 12, 525−536

newly synthesized proteins emerge within the next 30 min (T0.5). As suggested from experiments with transcriptional inhibitors (K. Mikulik, J. Bobek, personal communication), these earliest translated proteins might originate from mRNA transcription initiated during sporulation and the nascent messengers survive dormancy in stable polysomal complexes.

The earliest 30 min of germination can be assumed as an elemental recovery after breaking dormancy using inner energy sources prepared during sporulation. From T1 (30–60 min), we detected the rapid induction of proteosynthesis (see Figure 2). We presume that at this stage, the cells preserve fully competent active metabolism similar to responses to stress conditions.[21] During this period, cells detect energy sources present in the medium and adjust metabolic pathways through corresponding regulatory mechanisms that can be observed by the synthesis of pleiotropic regulators, such as BldD (SCO1489), Crp (SCO3571), transcription factor regulator SCO4232, anti-sigma factor Prs (SCO5244) and ribonuclease III (RNase III SCO5572), which are discussed below. The number of newly emerged proteins after the T1 stage decreased until T3.5 (180–210 min). Although the overall level of expression continued to increase, we did not detect any newly synthesized proteins after T3.5 (210 min).

Protein Functional Analysis

A total of 251 2D gel spots from germinating *S. coelicolor* were identified using MS (for a reference gel with identified spots see Supporting Information Figure 2). Several proteins were detected in more than one spot, presumably as a consequence of post-translational modifications or protein degradation. The resulting set of proteins contained 160 unique proteins, of which 114 proteins belonged to a set of *de novo* synthesized proteins expressed during the initial 5.5 h of germination.

These 160 identified proteins were distributed across 14 general functional groups (Figure 3) based on the functional

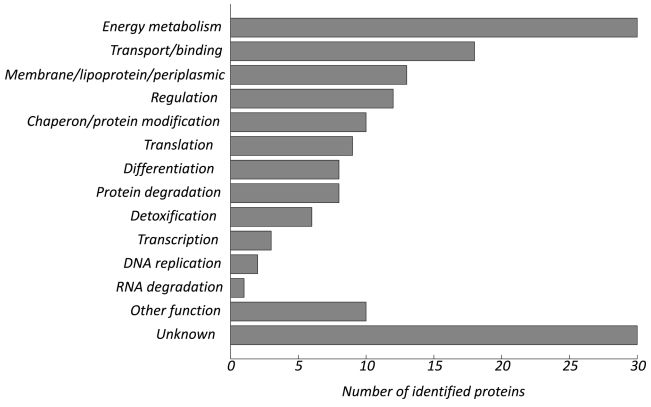


Figure 3. Functional classification of 160 identified proteins in germinating *S. coelicolor*.

(or predicted functional) classifications from the ftp://ftp.sanger.ac.uk/pub/S_coelicolor/classwise.txt and additional published information. The identified proteins were involved in energy metabolism (19%), transport/binding (11%), membrane/lipoprotein/periplasmic proteins (8%), regulation (7%), chaperones/protein modification (6%), translation (6%), differentiation (5%), protein degradation (5%), detoxification (4%), transcription (2%), DNA replication (1%) and RNA degradation (1%). Moreover, 6% of the proteins could not be assigned to any of the groups and were designated as proteins with other functions. Approximately 19% of the proteins had unknown functions.

Several identified proteins were present in dormant cells and might be required after their reactivation during germination. The expression of more than 70% of the proteins identified here was initiated during germination (radiolabeled). A list of detected proteins assigned to functional groups is presented in Supporting Information Table 3.

Protein Cluster Description

Protein Accumulation (SYPRO Ruby Staining). The resulting clusters of protein accumulation (SYPRO Ruby staining) are shown in Figure 4.

The proteins of clusters 1 and 2 were important for the initiation of germination, with increased concentrations observed at the first time points of the measurements. Altogether, these two clusters contain 37 identified protein spots. When assessing the frequency of the occurrence of individual functional groups and protein in the whole set, the only enriched group in this set was the "regulation" group (1.72-fold higher percentage) characterized by Crp, GntR family transcriptional regulator SCO5100, two-component response regulator SCO1801 and hypothetical protein SCO2127 (discussed below).

Interestingly, the profile of cluster 3 represented proteins in dormant spores showing decreased expression during the first time interval (T Dorm–T0) with no subsequent increase. The cluster was large, containing 88 spots, of which 57 were identified. The frequency of the functional groups inclines to proteins in the "membrane/lipoprotein/periplasmic" (2.3-fold higher percentage) and "transport/binding" (1.9-fold higher percentage) groups compared with the entire accumulation set. The "membrane/lipoprotein/periplasmic" group primarily comprised hypothetical proteins assigned to functional groups on a homology basis, and the "transport/binding" group included distinct proteins for the transport of small molecules. Therefore, it is doubtful that we will be able to draw conclusions from these data.

Cluster 4 (38 identified proteins from 78 spots) peaked at T1.5. A comparison of the percentage of proteins in the functional groups within the clusters and the entire accumulation set showed that enrichment could only be detected for proteins involved in "protein degradation" (2.2-fold).

The continual increase in protein accumulation was observed for spots in cluster 5 (46 identified proteins from 97 spots). Compared with the entire accumulation set, cluster 5 is enriched in protein from the functional group "chaperones/protein modification" (3.5-fold) and "translation" (2.5-fold). The accumulation of these proteins suggests a continuously increasing demand for proteosynthesis associated with the initiation of vegetative growth and increment biomass.

Clusters 6 and 7 contained too few identified proteins to draw any conclusions; these proteins are listed in the Supporting Information Table 3.

Protein Synthesis (Radiolabeling). The clusters were cleaned for the influence of the growth component expressed in the first eigenvector (see Materials and Methods, SVD) and indicated the *de novo* synthesis of proteins (Figure 5). The cluster profiles shown in Figure 5 can be characterized by constant proteosynthetic activity during the first two hours, which started to change after three hours of growth for cluster 6 and subsequently changed for all other clusters. Most of the

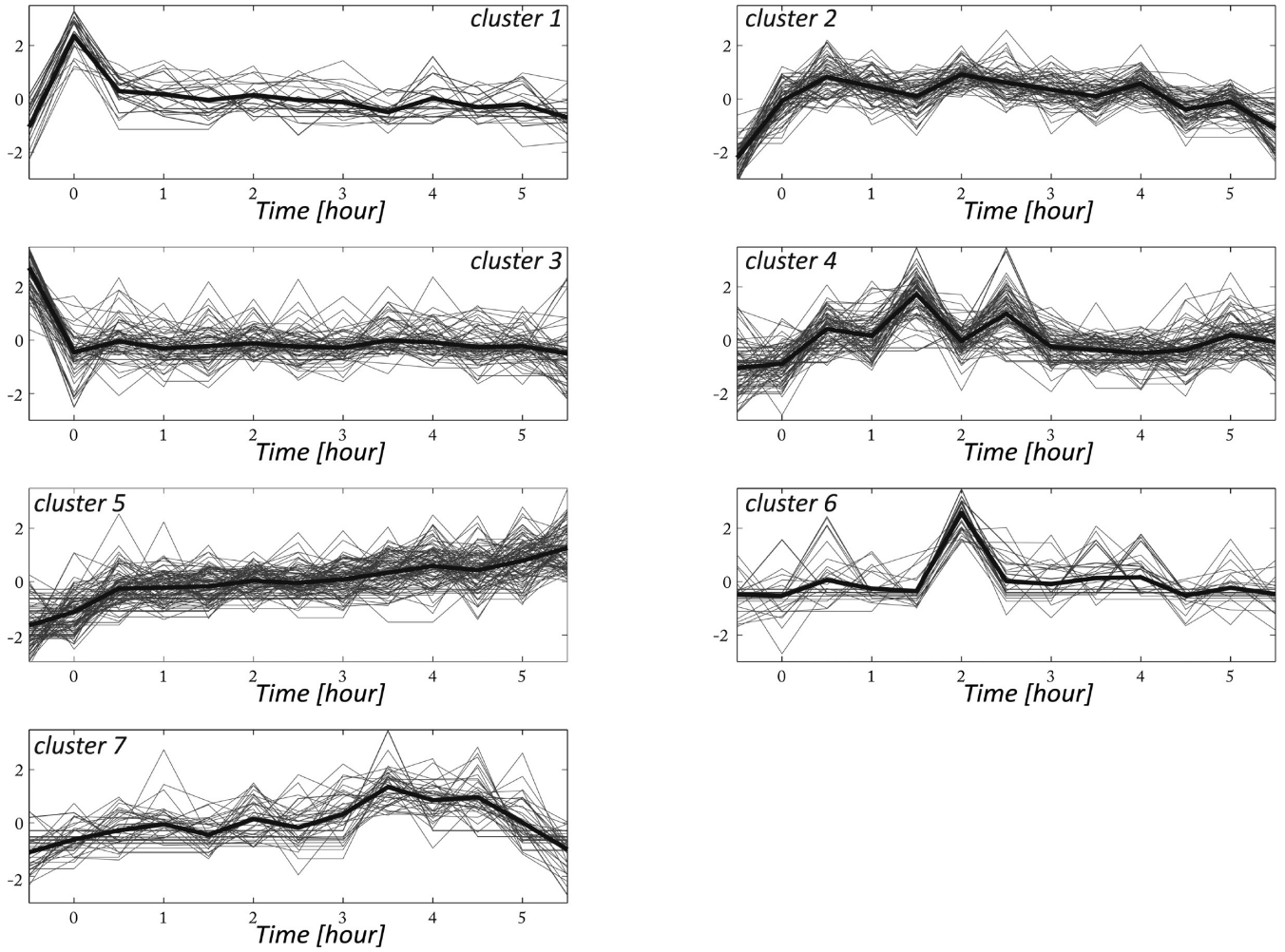


Figure 4. The hairlines represent protein profiles stained with SYPRO Ruby (protein accumulation) and grouped into 7 clusters. The thick line represents the cluster average. The profiles were standardized to dimensionless values.

proteosynthetic activity occurred in the last period (4–6 h), with peaks of expression varying over time for different clusters.

Cluster 1 peaked at time point T4.5 and contained 59 protein spots, of which 29 were identified. In comparison with the entire *de novo* synthesized set, this cluster was enriched in the "chaperones/protein modification" (2.3-fold higher percentage) and "transport/binding" (1.62-fold higher percentage) groups. A similar contour was obtained for proteins in cluster 2, increasing up to T4.5 (26 identified proteins from 71 protein spots), and proteins from functional groups "energy metabolism" (1.7-fold higher percentage) and "membrane/lipoprotein/periplasmic" (1.5-fold higher percentage) were enriched.

Clusters 3 and 4, peaking at T5, contained 55 proteins identified from 130 spots. A comparison of the percentage of proteins in the functional groups in the clusters and in the entire *de novo* synthesized set comprised the "regulation" group with percentage 1.5 times higher than for the entire set. More than half of all identified proteins from the "differentiation" group belonged to clusters 3 and 4, containing the cell division protein FtsZ (SCO2082) and cell growth-associated protein FilP (SCO5396) (proteins discussed below). The expression of proteins for the regulation of the onset of cell division and germ tube growth might define the transition from germination to active growth.

Cluster 5 (15 identified proteins from 52 protein spots), showed an increase at time T4 but did not show any specific enrichment of any functional groups. The number of identified proteins in cluster 6 was too small (5 proteins).

The complete membership of the proteins in the individual clusters, either for SYPRO Ruby staining or radiolabeling, is listed in Supporting Information Table 3.

Expression Profiles Correlation

SYPRO Ruby stained all proteins present in the extracted sample. After separation on 2D gels, the resulting fluorescent image represents the accumulated total amounts of proteins individually projected in the 2D gel spots. In contrast, radiolabeling only indicates the proteins *de novo* synthesized during the labeling interval. Protein spots matched over the course of fluorescent and radioactive gel sets designates a protein expression profile, indicating temporal changes in protein accumulation and synthesis, respectively.

As protein accumulation reflects the rates of synthesis and degradation, a comparison of SYPRO Ruby stained and radiolabeled protein expression profiles provides information about the dynamics of protein accumulation, including delays in protein synthesis from degradation and posttranslational modifications. To investigate these relationships, we correlated (Pearson correlation) the SYPRO Ruby-stained expression profiles of individual spots with the corresponding radiolabeled

dx.doi.org/10.1021/pr300980v | J. Proteome Res. 2013, 12, 525–536

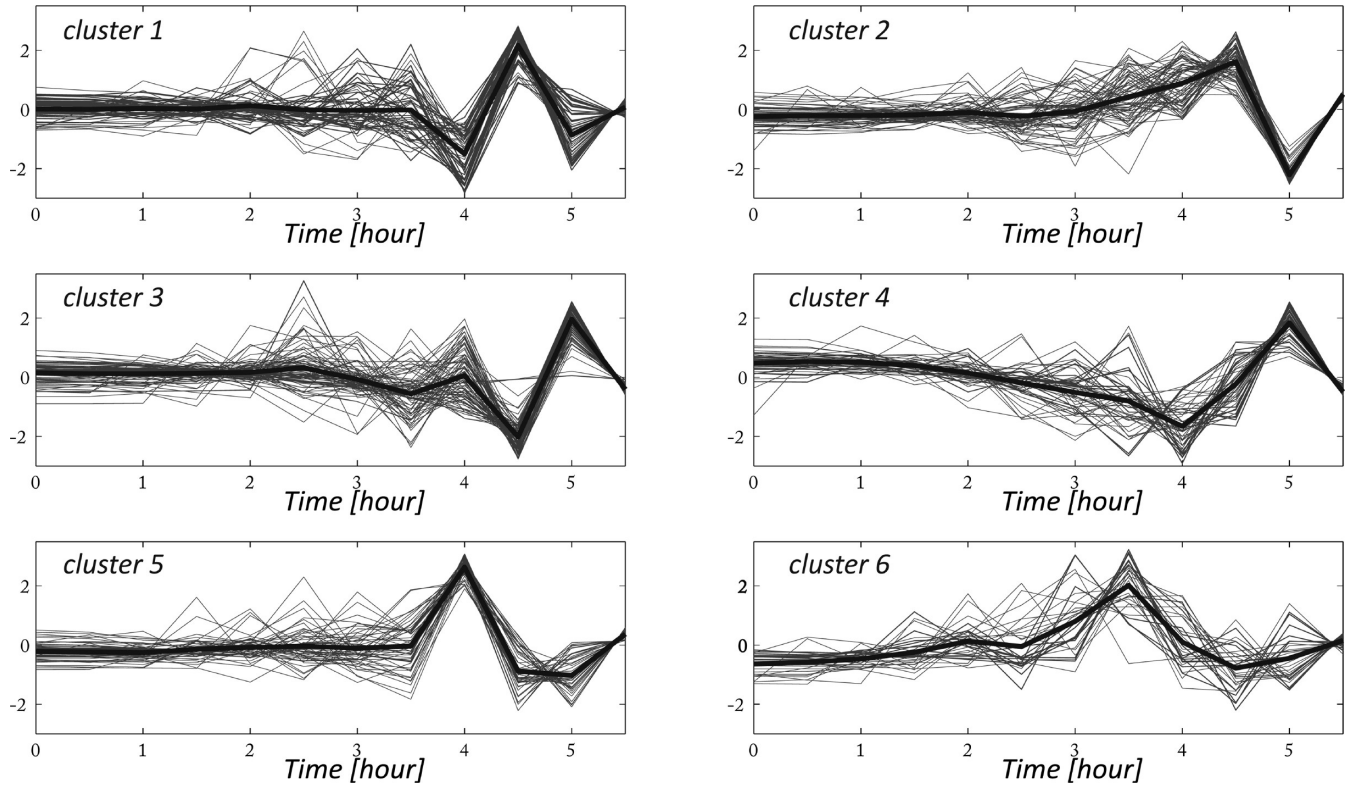


Figure 5. The hairlines represent radioactively labeled (proteins synthesis) protein profiles with the first eigenvector filtered out. The profiles were grouped into 6 clusters. The thick line indicates the cluster average. The profiles were standardized to dimensionless values.

spot profiles. The distribution of the Pearson correlation coefficient is shown in Figure 6. The histogram illustrates that

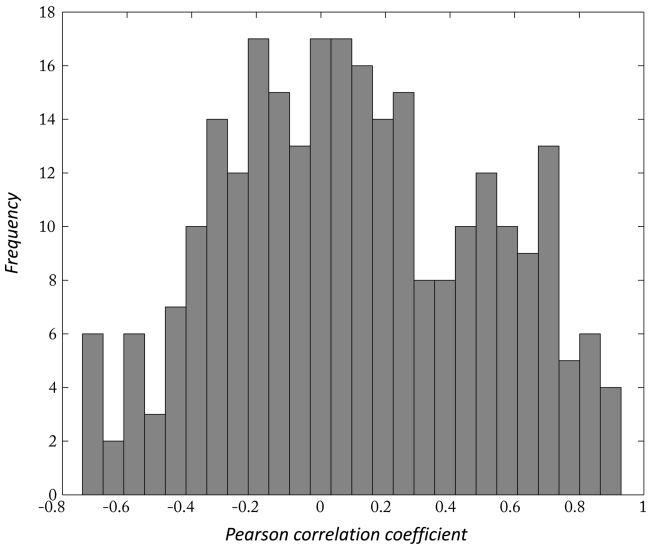


Figure 6. Histogram of the distribution of Pearson correlation coefficients among SYPRO Ruby stained and radiolabeled protein expression profiles.

the correlation coefficients exhibit two peaks: first around 0.1, second around 0.6. This result shows that the correlation between protein accumulation and synthesis is low or even negative for three-quarters of the protein profiles (74%). The remaining profiles, representing roughly 26%, depict protein profiles with mutually positive correlations.

Post-translational modifications or protein degradation presumably occurred when accumulation declined while protein synthesis increased, as observed for a majority of the expression profiles with negative correlations. If the synthesis was faster than the modification, then the volume of the corresponding SYPRO Ruby-stained electrophoretic spot was decreased, while the radiolabeled-spot continued to increase over time. Alternatively, an increased degradation of accumulated protein could account for these results. Most likely, the effect is a combination of both processes, suggesting that post-translational processes controlling protein function play a strong role in the control of gene expression. However, it remains unknown whether this phenomenon is pronounced during germination, which is a fundamental developmental change, or whether it can also be observed during the exponential phase of a steadily growing population. In any case, this observation is of considerable importance and should become a topic of an individual and more detailed study.

Analysis of Functional Groups in Identified Germination Proteins

To gain insight into the molecular processes of protein expression in the germinating spore, the initiation of protein expression was individually analyzed for the eight most occupied protein functional groups. The selection of the functional groups was based on the number of radiolabeled proteins within each functional group. Those groups having less than seven proteins within the group were excluded from the analysis. Figure 7 shows that the initiation of translation, which occurs immediately after the induction of germination (T0), generates the proteins required for translational machinery, differentiation and proteins acting as chaperones or proteins modifiers. Subsequently, the synthesis of new proteins is

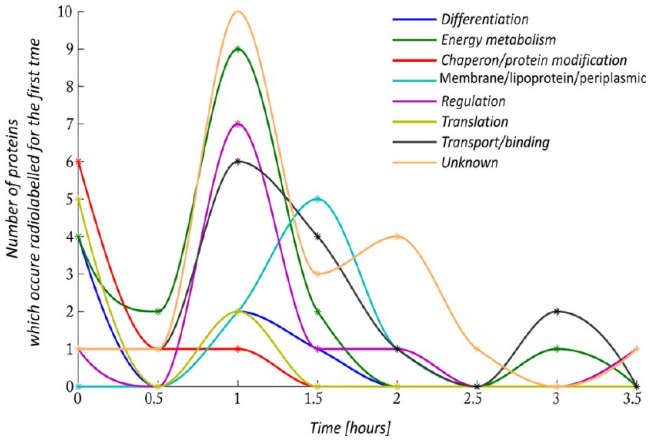


Figure 7. Progression in number of *de novo* synthesized proteins that initiate expression within the measured time periods for the 8 most occupied functional groups.

reduced (T0.5) and re-emerges at T1. The majority of the proteins first emerge at T1, where the expression of proteins involved in energy metabolism, regulation and transport/binding is initiated. According to Figures 2 and 7, the T1 phase represents a crucial moment in protein expression. Most of the proteins are newly synthesized at that time, including those with unknown function, suggesting their importance during the process of germination. Interestingly, all proteins classified as chaperones or involved in translation, which were detected in this study, were already expressed before point T1. These data confirm previously published results showing a role for chaperones expressed at an earlier stage of germination, which assist in the refolding of proteins aggregated during dormancy, the reactivation of inactive ribosomes and the correct folding of nascent proteins to their functional forms.[7] The increased expression of several heat shock proteins might also be induced during the initial 10 min cultivation at 50 °C used here to activate faster and synchronous spore germination. As shown in Figure 2, the number of specific proteins synthesized after breaking dormancy (black columns) declines to zero after 4 h, indicating that cells experienced the first step of differentiation within the first 3.5 h and successfully entered into vegetative growth.

Analysis of Functional Groups

Cell Wall Degradation and Energy Metabolism. The energy metabolism proteins comprised those involved in glycolysis, TCA cycle, ATP-proton motive force, fatty acid synthesis and hydrolases. The enzymes identified in this group are listed in Supporting Information Table 3.

To induce the biosynthesis leading to growth and morphological development, the spore must provide enzymes for energy metabolism. The distribution of the onset of the synthesis of these proteins over time (Figure 7) shows that the germinating spore primarily uses endogenous reserves as a source of energy. Catabolic processes degrade former spore compounds to initiate the rebuilding of the cell to the vegetative form. This process is associated with the release of energy. In the process of cell wall reconstruction, lysozyme-like enzymes mediate the lysis of peptidoglycan during germination. It was previously shown in *S. coelicolor* that the cell wall hydrolases are involved.[13] Here, we identified four hydrolases; SCO1061, SCO1725, SCO3487 and SCO5466. Interestingly, these enzymes were stored in dormant spores and none of them were *de novo* synthesized during germination, suggesting that cells prepare all required protein apparatuses during sporulation to restore active metabolism after the dormancy period. As our data show, SCO1061 and SCO1725 are abundant only in dormant spores and subsequently were not detected at all. SCO1725 was shown to cleave lipids.[22] The other two hydrolases, SCO3487 and SCO5466, exhibited almost constant expression during the observed 5.5 h period. The hydrolase SCO3487 accumulates during sporulation on solid agar plates due to its ability to degrade agar.[23] The peptidoglycan hydrolase SCO5466 is presumed to act as an N-acetylmuramidase/N-acetylglucosaminidase[13] and might be involved in the regulation of cell wall thickness.[12]

Chaperones/Protein Modification. As previously discussed, spore dormancy is characterized by low water content.[24] At this stage, most proteins are generated during sporulation, forming insoluble aggregates. To reactivate and initiate metabolic activity after the influx of water during spore activation, the aggregated proteins are refolded and/or modified by chaperones and protein modification enzymes.

Therefore, some chaperones were also stored in the dormant spore (GroES SCO4761 detected here as nonradioactive); however, the initiation of the expression of most chaperones was detected in this study during spore activation at T0. On one hand, the induction of chaperones might occur in response to heat shock (50 °C, 10 min), as many molecular chaperones are members of a family of heat shock proteins.[25] On the other hand, the constitutive expression of chaperones during the observed 5.5 h suggests their further requirement in germination, as they also assist in the folding of newly expressed proteins to avoid protein misfolding and/or aggregation. These chaperones are the trigger factor (SCO2620), GroEL1 (SCO4762), GroEL2 (SCO4296), DnaK (SCO3671), GrpE (SCO3670) and peptidyl-prolyl cis−trans isomerase (SCO1638, SCO7510, and SCO1639) that mediate protein interconversion. Several of these enzymes were detected prior to their expression during spore germination in *S. granaticolor*,[7] and were also shown to undergo reversible phosphorylation and bind active ribosomes to mediate the folding of nascent proteins.

Detoxification. Spores come in contact with radicals and toxic molecules that are natural products of active metabolism within the cell. Immediately during spore activation at time T0, the expression of two thioredoxins, TrxA (SCO3889) and TrxA4 (SCO5419), is initiated and both enzymes are involved in the antioxidant system, which protects against oxidative stress and maintains intracellular thiol homeostasis (reviewed in den Hengst and Buttner[26]).

Two representatives of the detoxification function group, superoxide dismutase sodF2 (SCO0999) and catalase SCO0379, which are functionally linked, were similarly detected at T1.5 of germination. First, superoxide dismutase (SodF2) catalyzes the conversion of superoxide anion to hydrogen peroxide and molecular oxygen. Superoxide dismutase is well documented enzyme in *S. coelicolor*.[27] Subsequently, catalase SCO0379 decomposes the toxic peroxide. The requirement for detoxification might reflect the increased activity of metabolism for up to 60 min of incubation (T1), when the expression of a majority of proteins has been initiated.

Regulation. The assumption that germination involves the transition from stress to directed expression is supported by the fact that the expression of most of the proteins classified here as

regulators is detected no earlier than T1. From T1 we have detected the increasing expression of Crp, which was suggested as a central regulatory protein in all developmental steps in *S. coelicolor*.[12] Piette et al. has shown that the expression of Crp is life cycle dependent and is expressed exclusively in the first part of the life cycle until the emergence of aerial hyphae. Deletion of *crp* strongly affects both germination and sporulation. However, the mechanism of Crp function is unknown, as explicit DNA targets for Crp have not been specified.

The proposed increasing importance of the regulation of gene expression during germination was also supported by the detection of the RNase III enzyme. Small non-protein-coding RNAs participate in gene expression regulation at the post-transcriptional level via blocking the translation of target mRNA through antisense binding. The RNase III enzyme subsequently degrades the double-stranded sRNA−mRNA. The importance of this enzyme in *Streptomyces* is demonstrated by the fact that RNase III null mutants do not produce antibiotics.[28]

Streptomyces possess astonishing numbers of transcriptional regulators to direct gene expression during their complex cell development. Our proteomic approach failed to detect the expression of sigma factors during spore germination. However, several aspects of the gene expression control in germination might be deduced from the detected expression of other regulators. We have shown the synthesis of an anti-sigma factor antagonist BldG (SCO3549) and a σ^H anti-sigma factor Prs (also called UshX) whose expression was initiated at T0 and T1, respectively. BldG and Prs were recently demonstrated to interact directly and a model of activation of σ^H (SCO5243, not detected here) through a partner-switching-like mechanism between BldG and Prs was proposed.[29] According to that model, a potential role for σ^H in early germination might be expected, based on the observation of the anti-sigma factor antagonist BldG and anti-sigma factor Prs. The σ^H has been suggested to regulate both the osmotic stress response and morphological differentiation,[30] indicating its importance during germination.

In addition to Prs, we detected in our germination time course from T1 expression of another regulator BldD, which also negatively controls the expression of the σ^H gene.[31] BldD is a small (18 kDa) DNA-binding protein whose pleiotropic activity is required for proper morphological differentiation and antibiotic production.[32] BldD acts primarily as a transcriptional repressor of developmental genes during vegetative growth. ChIP-chip experiments revealed that out of a total 167 binding sites on the chromosome, BldD controls the expression of other regulatory proteins, including those associated with signaling pathways, energy storage, proteolytic functions, cell division, cell wall modification, morphological differentiation and antibiotic production.[33]

The expression of SCO2127 detected here from T0.5 was previously shown to be involved in carbon catabolite repression of mycelium differentiation.[34] The authors showed that the SCO2127 interacts with the ABC transporter lipoprotein BldKB (SCO5113, detected here at T3). They suggested that the interaction between SCO2127 and BldKB might inhibit BldKB function as an oligopeptide transporter for aerial mycelium development. Thus, in the presence of glucose, SCO2127 might keep the cells in the vegetative phase of growth.

Differentiation. Interestingly, in our germination assay, we detected the expression of several proteins involved in cytoskeleton formation during growth and development, DivIVA (SCO2077), FilP and FtsZ.

The promoter for the gene encoding the cell division protein FtsZ is one of the BldD targets. FtsZ is a tubulin homologue, which during sporulation forms rings separating unique spores.[35] Its increasing expression, which starts at T1, suggests an additional role for this protein during germination.

Microscopically visible morphological changes in the germinating spore of *S. coelicolor* were the emergence of the germ tube. Under growth conditions in rich liquid AM medium (see Materials and Methods), the appearance of one or two germ tubes was observed at 3 h after the initiation of germination. For the formation of germ tubes, the presence of DivIVA[36,37] is indispensable. The coiled-coil protein DivIVA is an ortholog present in many bacteria, which serves as a marker for a site where tip extension and growth are established. DivIVA is a component of the machinery for cell wall biosynthesis. We observed that the *divIVA* gene is expressed immediately during the heat shock activation of spores at T0 and its synthesis increases with time. These results experimentally verify the idea that the expression of DivIVA is initiated early in germination.[36]

Another coiled-coil protein FilP might be associated with DivIVA.[38] FilP is characterized by a rod-shaped domain similar to intermediate filaments. Using green fluorescent protein fused to the C-terminus of FilP protein, it was shown that this protein is expressed in vegetative and aerial hyphae but not in mature spores.[38] FilP protein is thus associated with cell growth, strongly accumulating in the tips of young hyphae. Our data showed that this protein is expressed during the early stages of germination (T1.5). Hence, the first detected expression of FilP protein might be proposed as a marker characterizing the beginning of active growth.

Interestingly, during the activation of spores at T0, the expression of SapA (SCO0409) was observed. SapA belongs to the spore-associated protein (SAP) group, which is required for proper aerial mycelium formation and spore development.[39] The transcription of SapA occurs before the beginning of spore formation when the aerial mycelium appears.[39] Therefore, its expression during germination was unexpected. Although dormant spores were washed to remove mycelial debris, one might argue that the expression of these sporogenic proteins originates from contamination by spore-forming aerial hyphae. We do not expect contamination because the expression of SapA increases during the germination time course, indicating that the SapA protein is also required during spore germination.

Among proteins whose expressions begin early in germination are Tdd8 (SCO2368), SC4277 and SCO0641, which all contain a TerD domain. The function of proteins with a TerD domain is associated with tellurium resistance. Functional assignment is based on homology with the corresponding TerD protein present in *Serratia marcescens*, which is responsible for tellurium resistance.[40] However, a study[41] of *S. coelicolor* suggests that these proteins might possess a different function, as mutant strains with deletions and the overexpression of the *tdd8* gene exhibit a significant effect on *S. coelicolor* growth, differentiation and sporulation. Germinating spores express Tdd8 and SCO4277 during activation at T0, and their expression rapidly increases. The expression of SCO0641 begins at T1. Consistent with the results of previous studies,[41] the early production of Tdd8 and SCO4277 in germination might suggest their importance in *Streptomyces* development rather than in the tellurium resistance function. In addition,

Tdd8 and SCO4277 were experimentally detected among the most abundant proteins in the *S. coelicolor* proteome during the stationary phase,[2,42] and due to their high codon adaptation index (CAI), both proteins were predicted among highly expressed housekeeping genes associated with cell growth.[43]

CONCLUDING REMARKS

We have conducted a proteomic study, monitoring protein accumulation and synthesis. We analyzed protein expression during the germination of *S. coelicolor* as a principal developmental transition from the null dormant state to active metabolism.

Cell dormancy is an important phase of bacterial development characteristic not only for spore-forming bacteria. This stage might also be characterized by minimal metabolic activity. It is generally accepted that spores require nutrient sources to start the process of germination, as previously reviewed.[44] However, we presume that the pure water milieu is sufficient to interrupt spore dormancy. When spore germination was examined in distilled water used as cultivation medium lacking any nutrient sources, active proteosynthesis, as detected using 2D autoradiograms, was comparable to that from the standard AM medium (J. Bobek, unpublished observations). This result shows that dormant spores possess the compounds required for the successful initiation of active metabolism. It has also been suggested that the first phase of the active developmental program is independent of actual growth conditions. More likely it is predetermined genetically or by conditions occurring during sporulation.

We assume that during the activation of spore metabolism, the inner regulatory mechanisms of cells are "awakened" through liquid influx into the cytoplasm. The hydration of the spore cytoplasm restores aggregated proteins prepared in the cell during sporulation along with molecular chaperones that facilitate protein folding into their native forms.

We presume that the gene expression during germination (particularly the transition to T1) evokes the cell response to stress conditions. As shown in Figure 2, most proteins begin expression during the first hour of germination. During this period, cells launch basal metabolism to ensure the availability of all required metabolite intermediates, while detecting external signals and keeping active only those metabolic pathways actually required. This mechanism was also supported by the detected expression of several regulatory proteins and proteins essential for developmental differentiation at T1.

Another remarkable contribution of this study is a list of proteins detected during germination that have not been functionally characterized, particularly proteins synthesized during the onset of germination. Their expression profiles might serve as a data source for further investigations. The functional characterization of these "missing" proteins will provide deeper insight into the particular metabolic and regulatory pathways involved in the regulation of germination, which with the current state of knowledge, cannot be obtained.

ASSOCIATED CONTENT

Supporting Information

Supporting Information Table 1, normalized intensities of 2DE gel spots for both types of labeling; Supporting Information Table 2, MS protein identification details; Supporting Information Table 3, the list of identified proteins from germination of *S. coelicolor* and their ordering into functional groups and clusters; Supporting Information Figure 1, the reference 2DE gel image of proteins from germination of *S. coelicolor*; Supporting Information Figure 2, the reference 2DE gel image with marked identified proteins. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*Address: Institute of Microbiology, Academy of Sciences of the Czech Republic, Laboratory of Bioinformatics. E-mail: vohr@biomed.cas.cz.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported through grants P302-11-0229, P302/10/0468, and 310/07/1009 from the Czech Science Foundation, Grant Agency of the Charles University under contract no. 17409, Charles University grant no. SVV-2012-264506 and a long-term organization development plan no. 1011 from the Faculty of Military Health Sciences, University of Defence, Czech Republic. We would like to thank Dr. P. Stopka, Department of Zoology, Charles University in Prague for help with MS sample preparation. J.B. was supported by the project of Charles University in Prague: PRVOUK-P24/LF1/3.

REFERENCES

(1) Kieser, T.; Bibb, M. J.; Buttner, M. J.; Chater, K. F.; Hopwood, D. A. *Practical Streptomyces Genetics*; John Innes Foundation: Norwich, U.K., 2000.

(2) Jayapal, K. P.; Philp, R. J.; Kok, Y. J.; Yap, M. G.; Sherman, D. H.; Griffin, T. J.; Hu, W. S. Uncovering genes with divergent mRNA-protein dynamics in Streptomyces coelicolor. *PLoS One* **2008**, *3* (5), e2097.

(3) Manteca, A.; Sanchez, J.; Jung, H. R.; Schwammle, V.; Jensen, O. N. Quantitative proteomics analysis of Streptomyces coelicolor development demonstrates that onset of secondary metabolism coincides with hypha differentiation. *Mol. Cell. Proteomics* **2010**, *9* (7), 1423−1436.

(4) Vohradsky, J.; Branny, P.; Thompson, C. J. Comparative analysis of gene expression on mRNA and protein level during development of Streptomyces cultures by using singular value decomposition. *Proteomics* **2007**, *7* (21), 3853−3866.

(5) Nieselt, K.; Battke, F.; Herbig, A.; Bruheim, P.; Wentzel, A.; Jakobsen, O. M.; Sletta, H.; Alam, M. T.; Merlo, M. E.; Moore, J.; Omara, W. A.; Morrissey, E. R.; Juarez-Hermosillo, M. A.; Rodriguez-Garcia, A.; Nentwich, M.; Thomas, L.; Iqbal, M.; Legaie, R.; Gaze, W. H.; Challis, G. L.; Jansen, R. C.; Dijkhuizen, L.; Rand, D. A.; Wild, D. L.; Bonin, M.; Reuther, J.; Wohlleben, W.; Smith, M. C.; Burroughs, N. J.; Martin, J. F.; Hodgson, D. A.; Takano, E.; Breitling, R.; Ellingsen, T. E.; Wellington, E. M. The dynamic architecture of the metabolic switch in Streptomyces coelicolor. *BMC Genomics* **2010**, *11*, 10.

(6) Thomas, L.; Hodgson, D. A.; Wentzel, A.; Nieselt, K.; Ellingsen, T. E.; Moore, J.; Morrissey, E. R.; Legaie, R.; Wohlleben, W.; Rodriguez-Garcia, A.; Martin, J. F.; Burroughs, N. J.; Wellington, E. M.; Smith, M. C. Metabolic switches and adaptations deduced from the proteomes of Streptomyces coelicolor wild type and phoP mutant grown in batch culture. *Mol. Cell. Proteomics* **2011**, *11* (2), M111 013797.

(7) Bobek, J.; Halada, P.; Angelis, J.; Vohradsky, J.; Mikulik, K. Activation and expression of proteins during synchronous germination of aerial spores of Streptomyces granaticolor. *Proteomics* **2004**, *4* (12), 3864−3880.

(8) Mikulik, K.; Bobek, J.; Bezouskova, S.; Benada, O.; Kofronova, O. Expression of proteins and protein kinase activity during germination

of aerial spores of Streptomyces granaticolor. *Biochem. Biophys. Res. Commun.* **2002**, *299* (2), 335−342.
(9) Susstrunk, U.; Pidoux, J.; Taubert, S.; Ullmann, A.; Thompson, C. J. Pleiotropic effects of cAMP on germination, antibiotic biosynthesis and morphological development in Streptomyces coelicolor. *Mol. Microbiol.* **1998**, *30* (1), 33−46.
(10) Derouaux, A.; Dehareng, D.; Lecocq, E.; Halici, S.; Nothaft, H.; Giannotta, F.; Moutzourelis, G.; Dusart, J.; Devreese, B.; Titgemeyer, F.; Van Beeumen, J.; Rigali, S. Crp of Streptomyces coelicolor is the third transcription factor of the large CRP-FNR superfamily able to bind cAMP. *Biochem. Biophys. Res. Commun.* **2004**, *325* (3), 983−990.
(11) Derouaux, A.; Halici, S.; Nothaft, H.; Neutelings, T.; Moutzourelis, G.; Dusart, J.; Titgemeyer, F.; Rigali, S. Deletion of a cyclic AMP receptor protein homologue diminishes germination and affects morphological development of Streptomyces coelicolor. *J. Bacteriol.* **2004**, *186* (6), 1893−1897.
(12) Piette, A.; Derouaux, A.; Gerkens, P.; Noens, E. E.; Mazzucchelli, G.; Vion, S.; Koerten, H. K.; Titgemeyer, F.; De Pauw, E.; Leprince, P.; van Wezel, G. P.; Galleni, M.; Rigali, S. From dormant to germinating spores of Streptomyces coelicolor A3(2): new perspectives from the crp null mutant. *J. Proteome Res.* **2005**, *4* (5), 1699−1708.
(13) Haiser, H. J.; Yousef, M. R.; Elliot, M. A. Cell wall hydrolases affect germination, vegetative growth, and sporulation in Streptomyces coelicolor. *J. Bacteriol.* **2009**, *191* (21), 6501−6512.
(14) Garrels, J. I.; Franza, B. R.; Chang, C.; Latter, G. Quantitative exploration of the REF52 protein database: cluster analysis reveals the major protein expression profiles in responses to growth regulation, serum stimulation, and viral transformation. *Electrophoresis* **1990**, *11* (12), 1114−1130.
(15) Molloy, M. P.; Brzezinski, E. E.; Hang, J.; McDowell, M. T.; VanBogelen, R. A. Overcoming technical variation and biological variation in quantitative proteomics. *Proteomics* **2003**, *3* (10), 1912−1919.
(16) Shevchenko, A.; Tomas, H.; Havlis, J.; Olsen, J. V.; Mann, M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat. Protoc.* **2006**, *1* (6), 2856−2860.
(17) Chamrad, I.; Strouhal, O.; Rehulka, P.; Lenobel, R.; Sebela, M. Microscale affinity purification of trypsin reduces background peptides in matrix-assisted laser desorption/ionization mass spectrometry of protein digests. *J. Proteomics* **2011**, *74* (7), 948−957.
(18) Rappsilber, J.; Mann, M.; Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2007**, *2* (8), 1896−1906.
(19) Rehulkova, H.; Chalupova, J.; Sebela, M.; Rehulka, P. A convenient purification and preconcentration of peptides with alpha-cyano-4-hydroxycinnamic acid matrix crystals in a pipette tip for matrix-assisted laser desorption/ionization mass spectrometry. *J. Mass Spectrom.* **2009**, *45* (1), 104−111.
(20) Sanders, S. K.; Alexander, E. L.; Braylan, R. C. A high-yield technique for preparing cells fixed in suspension for scanning electron microscopy. *J. Cell Biol.* **1975**, *67* (2 Pt.1), 476−480.
(21) Vohradsky, J.; Ramsden, J. J. Genome resource utilization during prokaryotic development. *FASEB J.* **2001**, *15* (11), 2054−2056.
(22) Cote, A.; Shareck, F. Cloning, purification and characterization of two lipases from Streptomyces coelicolor A3(2). *Enzyme Microb. Technol.* **2008**, *42*, 381−388.
(23) Temuujin, U.; Chi, W. J.; Chang, Y. K.; Hong, S. K. Identification and biochemical characterization of Sco3487 from Streptomyces coelicolor A3(2), an exo- and endo-type beta-agarase-producing neoagarobiose. *J. Bacteriol.* **2011**, *194* (1), 142−149.
(24) Ensign, J. C. Formation, properties, and germination of actinomycete spores. *Annu. Rev. Microbiol.* **1978**, *32*, 185−219.
(25) Gething, M. J.; Sambrook, J. Protein folding in the cell. *Nature* **1992**, *355* (6355), 33−45.
(26) den Hengst, C. D.; Buttner, M. J. Redox control in actinobacteria. *Biochim. Biophys. Acta* **2008**, *1780* (11), 1201−1216.
(27) Kim, E. J.; Chung, H. J.; Suh, B.; Hah, Y. C.; Roe, J. H. Expression and regulation of the sodF gene encoding iron- and zinc-containing superoxide dismutase in Streptomyces coelicolor Muller. *J. Bacteriol.* **1998**, *180* (8), 2014−2020.
(28) Adamidis, T.; Champness, W. Genetic analysis of absB, a Streptomyces coelicolor locus involved in global antibiotic regulation. *J. Bacteriol.* **1992**, *174* (14), 4622−4628.
(29) Sevcikova, B.; Rezuchova, B.; Homerova, D.; Kormanec, J. The anti-anti-sigma factor BldG is involved in activation of the stress response sigma factor sigma(H) in Streptomyces coelicolor A3(2). *J. Bacteriol.* **2010**, *192* (21), 5674−5681.
(30) Sevcikova, B.; Benada, O.; Kofronova, O.; Kormanec, J. Stress-response sigma factor sigma(H) is essential for morphological differentiation of Streptomyces coelicolor A3(2). *Arch. Microbiol.* **2001**, *177* (1), 98−106.
(31) Kelemen, G. H.; Viollier, P. H.; Tenor, J.; Marri, L.; Buttner, M. J.; Thompson, C. J. A connection between stress and development in the multicellular prokaryote Streptomyces coelicolor A3(2). *Mol. Microbiol.* **2001**, *40* (4), 804−814.
(32) Elliot, M.; Damji, F.; Passantino, R.; Chater, K.; Leskiw, B. The bldD gene of Streptomyces coelicolor A3(2): a regulatory gene involved in morphogenesis and antibiotic production. *J. Bacteriol.* **1998**, *180* (6), 1549−55.
(33) den Hengst, C. D.; Tran, N. T.; Bibb, M. J.; Chandra, G.; Leskiw, B. K.; Buttner, M. J. Genes essential for morphological development and antibiotic production in Streptomyces coelicolor are targets of BldD during vegetative growth. *Mol. Microbiol.* **2010**, *78* (2), 361−379.
(34) Chavez, A.; Forero, A.; Sanchez, M.; Rodriguez-Sanoja, R.; Mendoza-Hernandez, G.; Servin-Gonzalez, L.; Sanchez, B.; Garcia-Huante, Y.; Rocha, D.; Langley, E.; Ruiz, B.; Sanchez, S. Interaction of SCO2127 with BldKB and its possible connection to carbon catabolite regulation of morphological differentiation in Streptomyces coelicolor. *Appl. Microbiol. Biotechnol.* **2011**, *89* (3), 799−806.
(35) Grantcharova, N.; Lustig, U.; Flardh, K. Dynamics of FtsZ assembly during sporulation in Streptomyces coelicolor A3(2). *J. Bacteriol.* **2005**, *187* (9), 3227−3237.
(36) Flardh, K. Essential role of DivIVA in polar growth and morphogenesis in Streptomyces coelicolor A3(2). *Mol. Microbiol.* **2003**, *49* (6), 1523−1536.
(37) Flardh, K. Growth polarity and cell division in Streptomyces. *Curr. Opin. Microbiol.* **2003**, *6* (6), 564−571.
(38) Bagchi, S.; Tomenius, H.; Belova, L. M.; Ausmees, N. Intermediate filament-like proteins in bacteria and a cytoskeletal function in Streptomyces. *Mol. Microbiol.* **2008**, *70* (4), 1037−1050.
(39) Guijarro, J.; Santamaria, R.; Schauer, A.; Losick, R. Promoter determining the timing and spatial localization of transcription of a cloned Streptomyces coelicolor gene encoding a spore-associated polypeptide. *J. Bacteriol.* **1988**, *170* (4), 1895−1901.
(40) Whelan, K. F.; Colleran, E.; Taylor, D. E. Phage inhibition, colicin resistance, and tellurite resistance are encoded by a single cluster of genes on the IncHI2 plasmid R478. *J. Bacteriol.* **1995**, *177* (17), 5016−5027.
(41) Sanssouci, E.; Lerat, S.; Grondin, G.; Shareck, F.; Beaulieu, C. tdd8: a TerD domain-encoding gene involved in Streptomyces coelicolor differentiation. *Antonie van Leeuwenhoek* **2011**, *100* (3), 385−398.
(42) Langlois, P.; Bourassa, S.; Poirier, G. G.; Beaulieu, C. Identification of Streptomyces coelicolor proteins that are differentially expressed in the presence of plant material. *Appl. Environ. Microbiol.* **2003**, *69* (4), 1884−1889.
(43) Wu, G.; Culley, D. E.; Zhang, W. Predicted highly expressed genes in the genomes of Streptomyces coelicolor and Streptomyces avermitilis and the implications for their metabolism. *Microbiology* **2005**, *151* (Pt 7), 2175−2187.
(44) Flardh, K.; Buttner, M. J. Streptomyces morphogenetics: dissecting differentiation in a filamentous bacterium. *Nat. Rev. Microbiol.* **2009**, *7* (1), 36−49.

Appendix C

Paper III

Global Features of Gene Expression on the Proteome and Transcriptome Levels in *S. coelicolor* during Germination

Strakova, E., Bobek, J., Zikova, A. and Vohradsky, J.

PLoS ONE, accepted July 2013

GLOBAL FEATURES OF GENE EXPRESSION ON THE PROTEOME AND TRANSCRIPTOME LEVELS IN *S. COELICOLOR* DURING GERMINATION

Eva Strakova[1], Jan Bobek[1,2], Alice Zikova[1], Jiri Vohradsky[1]*

[1] Institute of Microbiology, Academy of Sciences of the Czech Republic, Laboratory of Bioinformatics, Vídeňská 1083, 142 20 Prague 4, Czech Republic.

[2] Institute of Immunology and Microbiology, First Faculty of Medicine, Charles University in Prague, Studničkova 7, 128 00 Praha 2, Czech Republic.

*corresponding author

e- mail: vohr@biomed.cas.cz

Abstract

Streptomycetes have been studied mostly as producers of secondary metabolites, while the transition from dormant spores to an exponentially growing culture has largely been ignored. Here, we focus on a comparative analysis of fluorescently and radioactively labeled proteome and microarray acquired transcriptome expressed during the germination of *Streptomyces coelicolor*. The time-dynamics is considered, starting from dormant spores through 5.5 hours of growth with 13 time points. Time series of the gene expressions were analyzed using correlation, principal components analysis and an analysis of coding genes utilization. Principal component analysis was used to identify principal kinetic trends in gene expression and the corresponding genes driving *S. coelicolor* germination. In contrast with the correlation analysis, global trends in the gene/protein expression reflected by the first principal components showed that the prominent patterns in both the protein and the mRNA domains are surprisingly well correlated. Analysis of the number of expressed genes identified functional groups activated during different time intervals of the germination.

1 Introduction

Bacterial cell dormancy is a calmness stage of living cells that is characterized by minimal metabolic activity. In several bacteria, the beginning of the dormancy is accompanied by a transition into a morphologically and physiologically distinct form, which is known as a dormant spore. The spore formation ensures that the cell will survive under unfavorable conditions. The transition process is called sporulation and has been well studied, not only in the species of *Bacillus* and *Clostridium* but also in *Streptomyces*. *Streptomyces* are Gram-positive bacteria that undergo a complex cell cycle that involves morphologically distinguishable developmental stages. The stages include unicellular spores that develop to branching substrate mycelium, which gives rise to the apically growing sporangium known as aerial mycelium, from which the spores are formed. Although spore germination is usually accepted as the first stage of the cell cycle, from the one-cell point of view, it represents the middle part of the life of a single cell, which starts during aerial mycelium formation and finishes in the developed substrate mycelium. The reverse process, awakening the cell back into a metabolically active form, is called germination and, as opposed to the sporulation, is much less understood in *Streptomyces*. In this multicellular bacterium, the dormant spores are the only haploid unicellular state. The spores are surrounded by a coat that protects the cellular content from an environmental challenge and enables survival. The spores also possess intracellular nutrient and energy sources, such as trehalose [1] or most likely polyphosphates (volutin) [2]. From the sporulation phase, dormant spores are also accommodated with a protein apparatus, such as chaperones and cell wall hydrolases, which are effective during metabolism renewal [3]. It has previously been suggested that dormant spores are devoid of stable functional mRNA [4]. However, further experiments, using rifampicin as a transcription initiation inhibitor, have revealed that dormant spores possess a pool of mRNAs, which provide templates for early protein synthesis [5]. All of these sources that arise from sporulation are thought to ease cell survival by triggering energy metabolism, which starts when the dormancy is broken and lasts until the cells are adapted to exploit external nutrient supplies.

In a favorable milieu, spores lose hydrophobicity, which allows water influx, which, in turn, initiates germination. Cells re-activate their metabolism and develop

into vegetative forms, building branching hyphae. Although several chemical or physical factors have been described for inducing germination, the molecular machinery that triggers the process remains unknown. Germination implies massive proteome reconstitution. This step is achieved by the utilization of spore compounds that are degraded during catabolic processes. Aggregated spore proteins that are preserved from dormancy are hydrated and re-activated during germination [6]. Cell wall hydrolases, such as RpfA and SwlA, provide the lysis of spore peptidoglycan to allow the entrance of external nutrients [7]. The re-activated chaperones GroEL, Trigger factor and DnaK were detected as assisting the reactivation of the proteosynthetic apparatus, which is fully accelerated during the first initial steps [8]. A systematic proteomic study that was recently conducted on streptomycete spore germination [3] revealed that the first newly expressed proteins are members of proteosynthetic machinery, proteins that are involved in differentiation, protein modifiers and other chaperones. Further protein expression evokes cellular responses to stress conditions and lasts approximately 1 hour. After this period, several protein regulators appear to take control over energy metabolism and further development. At this stage, the cell can respond to environmental conditions by direct gene expression.

Subsequent development to vegetative forms is a sequential process that is associated with the first DNA replication and an enhancement in the rate of RNA and protein synthesis [9]. Microscopically distinguishable germination tubes rise from the inner wall of spores and progress through the outer wall. The not-yet-fully defined end of the germination process is represented by the further protraction of the tubes, with the emergence of proteins that assist with cytoskeleton formation, such as DivIVA, FilP and FtsZ [3].

The correlation of the protein and mRNA expression levels has been most comprehensively reviewed by Abreu et al. [10], who summarize the knowledge of large-scale measurements on the level of whole transcriptomes and proteomes. Using 22 datasets that range from *E. coli* to Human, the authors document the correlation between protein and mRNA abundances, showing that a correlation coefficient (Pearson) ranges from 0.36 for *D. melanogaster* to 0.74 for *S. cerevisiae*. The authors also mention the importance of the quality of the measured data, which differs substantially between the protein and the mRNA measurements and is much lower for the proteomic data, especially in terms of the number of identified proteins in

comparison with the mRNAs. For an exact comparison of the transcriptome and proteome expression, improvements in the proteome quantification are essential. A comparison of the time series of the gene expression of both types of data has much less representation in the literature than comparisons of one-point measurements. In streptomycetes, only two papers have addressed this issue, with the first comparing transcriptomic and proteomic time series measured during exponential growth of *S. coelicolor* [11] and the second using the same species but focusing on the late exponential and stationary phases [12]. Both of these papers used a similar information-extraction method, which was singular value decomposition (SVD) [11] or principal components analysis (PCA) [12]. Several other papers on this topic, mostly concerning yeast, were published over the past decade [13,14,15,16,17]. Although an absolute correlation between the protein and the mRNA was on the same scale as reported for other species (r=0.63) [12], a striking similarity was found between the first PC loadings (eigenvectors), which were well correlated for the highest loadings, thus indicating a similarity in the expression between the protein and the mRNA of the backbone processes [13].

In this paper, we focus on the simultaneous analysis of gene and protein expression in *S. coelicolor* during the first 5.5 hours of germination, starting from dormant spores and sampled in 30-minute intervals. The obtained time series were compared using PCA and a correlation of the PC loadings with a time series of genes of the metabolic and regulatory functional groups.

The utilization of coding genes, i.e., how many genes and in what amounts they are expressed at individual time points, was performed on a symbolic level using generalized canonical law [18,19]. Here, we focus on identifying the absolute numbers of expressed genes as a function of growth, an approach that allowed us to trace how different functional groups of genes are expressed in the course of germination.

2 Materials and methods

2.1 Spore cultivation

S. coelicolor A3(2) M145 spores were pre-germinated in 2X YT media for 24 h (160 rpm, 30°C) [20]. Three milliliters of inoculum was transferred to solid agar plates (0.4% yeast extract, 1% malt extract, 0.4% glucose, 2.5% bacterial agar, pH

7.2) overlaid with cellophane discs and was cultivated for 14 days at 28°C. The harvested spores were used for germination in liquid AM medium. To boost the synchrony within the population, the spores were subjected to a 10-minutes heat shock treatment at 50°C. The protein and mRNA samples were collected at 30-min intervals starting from dormant spores until 5.5 hours of growth, obtaining samples at 13 time points. The phenotypic change occurring during germination is illustrated in the electron microscopy images of *S. coelicolor* spores (Supplementary Figure 1. A, B) for the dormant spores (T Dorm, Supplementary Figure 1. A) and for the germinating spores, 5,5 hours after germination initiation with grown germ tubes (Supplementary Figure 1. B).

2.2 Proteomics

Details concerning the sample preparation, 2D electrophoresis, radio labeling, fluorescent staining and MS identification of protein spots can be found in our previous publication [3]. Here, we mention only the steps that are essential for this paper.

Germinating spores were radiolabeled with ^{35}S Cysteine-methionine in 30-min radioactive pulses, except for the first time point (T0), which was labeled for 10 min during heat shock. Isolated protein samples were resolved by 2D gel electrophoresis using 24-cm strips with a pH range of 4-7. The second dimension was run on 12.5% polyacrylamide gels that were 25.5 x 20.5 cm in size, covering a Mw range of approximately 15-110 kDa. The gels were stained overnight with Sypro Ruby fluorescent dye and scanned on BioRad Phosphoimager FX for fluorescence intensity. Dried gels were exposed for 4 days to BAS cassettes (Fujifilm) and the protein radioactivity was determined using BioRad Phosphoimager FX. The stained and radioactive gel images were processed and compared using the software PDQuest 8.0.1 (Bio-Rad) to detect changes in the intensities for specific gel spots (proteins) over time and across replicates. Altogether, 54 2DE gels for Sypro Ruby staining and 50 radiolabeled 2DE gels were analyzed. Sypro Ruby-stained gels and radiolabeled gels were arranged into individual matchsets. The Sypro Ruby matchset reference gel contained 671 individual protein spots, and the radiolabeled matchset reference gel contained 404 spots. All of the gels were assembled into a single high-level matchset whose reference gel contained a total of 782 protein spots. All of the

visible spots were picked from a preparative gel and were analyzed by mass spectrometry. Details about MS identification and a complete list of characteristics of MS spectra are given in the supplementary materials of our preceding paper [3]. The experiment was designed to cover both the experimental and biological variance, combining the measurements from different technical and biological replicates at one time point. The numbers of 2DE gels that were used for the replicates in different time points are given in Table 1.

2.2.1 Proteomic data normalization

The 2D electrophoretic spot intensities in individual gels were standardized by dividing the spot intensities by the total protein concentration loaded on a gel. The multiplicative factor was calculated from the accumulative gels (Sypro Ruby staining). We assumed that the logarithm of the intensities on the accumulative gels was normally distributed, with the means distributed around a common mean. Therefore, the means of all of the spot distributions for all of the stained gels were averaged, and a multiplicative factor that adjusted all of the distributions to the same mean was computed for each gel. Because a single gel contained both the fluorescent and radioactive labeling, the multiplicative factors derived for the Sypro Ruby-stained gels were also used to normalize the radioactive-based images (for details see our previous work [3]). This method also enables correct normalization for the radiolabeled gels, for which only a few electrophoretic spots could be identified in the first time points. Using a cumulative radioactive signal for their normalization would lead to an inadequate amplification of the first time points in the electrophoretograms.

To assess the degree of variance in the quantification of the relative protein abundance levels, we calculated the coefficient of variation (CV) for replicates within the individual time points for both types of labeling. The values were computed from the normalized data. The mean CV for Sypro Ruby stained gels was 0.39, and it was 0.54 for radiolabeled gels. These values were distorted by a time dependence; the highest CV was observed for the earliest gels and decreased over time (CV_{max}=0.45, CV_{min}=0.27 for the stained gels; CV_{max}=0.62, CV_{min}=0.4 for the radiolabeled gels). These values were comparable with those previously reported in the literature for 2D gel electrophoresis experiments (20-40%) [21] and were higher

for the radiolabeled gels. The higher CV for the radiolabeled gels was caused mainly by the gels of the first time points, i.e., when proteosynthesis starts. The degree of variation is high both for technical reasons, when only a small number of spots appeared on a gel, and for the inherent variation among biological replicates, which is known to be high in *Streptomyces* in general. We attempted to overcome this problem by increasing the number of gels and sample replicates, which is almost two-fold higher than in a usual proteomic experiment of this scale (Table 1).

2.3 Transcriptomics

2.3.1 RNA isolation from spores

To break the cells, we used a FastPrep-24 machine (Biomedicals) in which the spores were mechanically disrupted in tubes containing zirconium sand, two 4-mm glass beads, and 500 µl of lysis buffer [22] (50 mM Tris-HCl pH8, 500 mM LiCl, 50 mM EDTA pH8, 5% SDS) and 8 µl of RNAse inhibitors (Biorad). The disruption was made in 6 rounds for 35 s, while the tubes were re-chilled between each round. The samples were centrifugated at 14000 g for 15 min at 4°C, and the supernatant was used to phenol-chloroform RNA extraction, which was repeated twice. The RNA precipitated overnight in ethanol and 3 M Sodium Acetate at -20°C. Finally, the RNA was re-suspended in 50 µl RNAse-free water and 0.5 µl RNAse inhibitors and was cleaned from possible DNA remains using the DNAse-Free kit (Ambion). The RNA was stored in water at -20°C.

2.3.2 DNA microarrays and data processing

The number of analyzed microarrays that represent 3 (or 2) biological and/or technical replicates for each experimental time point are given in Table 1. RNA quality control and gene expression levels were determined by Oxford Gene Technology (Oxford, UK) on Agilent DNA microarrays, covering the entire *S. coelicolor* genome, using OGT's standard Bacterial RNA amplification Protocol for the two-channel essay.

The acquired data were linear LOWESS normalized and filtered for background and flag information (from Agilent documentation) in the GeneSpring software to obtain genes that were significantly expressed above the background and to avoid the

side effects of possible cross-hybridizations. This step reduced the number of entities on a single array from 43888 to 25312, which represented the outcome for 7115 genes of the original 7825. The data discussed have been deposited in NCBI's Gene Expression Omnibus [23] and are accessible through GEO Series accession number GSE44415 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc= GSE44415).

2.3.3 Array normalization

The experiment included 37 arrays from 13 distinct time points of *S. coelicolor* germination. The arrays shared a common reference in the red channel (Cy5, beta channel), which was a mixture of RNA samples from all of the examined time points; the sample signal was recorded in the Cy3-labeled channel (alpha channel). The distributions of the Log2Ratio values (Log2Ratio= log2 (Sample (Cy3) / Reference (Cy5))) from each array were centered to ensure that the medians and the median absolute deviations of all array distributions were equal. The centering was performed by subtracting the Log2Ratio median value of the array from each Log2Ratio converted measurement on the array and dividing it by the median absolute deviation. To eliminate array outliers, we filtered out the 0.02 quantile of the least and most intensive Log2Ratio values. The normalized Log2Ratios were returned to the original scale by exponentiation (creating normalized Ratios).

The time series of the relative mRNA concentration was obtained by averaging the normalized Ratios across biological and technical replicates at specific time points and across all of the gene replicate spots that were presented on the array. Before averaging, the outliers among the gene replicates at one time point were filtered using the Q-test (for 3-9 inputs) and the Pierce test (for > 10 inputs).

The filtering caused the result that, in a few of the profiles, there was no value for certain time points. Such zero values were examined to determine whether they were placed between two non-zero time points. If the neighboring time points were non-zero, the missing value was linearly interpolated (which was performed for approximately 100 profiles of the total 7115). After filtering, the log2Ratio values were exponentiated to obtain the signal in its original scale. These measurements were arranged for individual genes into time series that form a “gene expression profile”, a term that will be used throughout this paper.

We further filtered out genes whose overall expression during germination was too low to be considered. The idea was to filter out genes whose microarray signal could be a result of array errors that were above the simple technical criteria. Therefore, a median expression level of the expression profile for all time points and biological and microarray replicates were computed. The expression level was defined as a raw signal from an individual chip spot, from the channel that recorded the fluorescently labeled mRNA sample (alpha channel). The median represents an overall expression level of individual genes. A logarithm base 2 of the expression profile medians was computed. The distribution followed a roughly lognormal shape, which indicated that there was a gap approximately at the position of the first quartile (data not shown). Therefore, we filtered out all of the genes whose log2median expression level was below the first quartile value. The genes that exhibit a single peak in the profile that would be otherwise filtered out were identified individually and were added to the final set. The final set contained 5385 genes. Among the removed genes, those that prevailed were the genes that have an unknown function and are unclassified and those that are associated with a secondary metabolism and are not expected to be expressed during germination.

2.4 Data treatment common to both proteomic and transcriptomic experiments

2.4.1 Standardization

In the PCA analysis, we considered the changes in the gene/protein expression patterns rather than in the absolute values. Therefore, all of the profiles from all experiments were normalized to have the same mean and variance by subtracting the mean of the individual profile from each member point of the profile and dividing it by the standard deviation of the profile.

2.4.2 Correlation

The Pearson correlation coefficient calculated below was tested under the assumption that each p-value is the probability of obtaining a correlation that is as large as the observed value by random chance, when the true correlation is zero. The p-value was computed using a t-statistic. If the p-value was ≤ 0.05, then the test was considered to be significant.

2.4.3 Chi-square test

Chi-square statistic was used to compare the number of gene products in different functional groups in any of the selected sets (any of the sets that were selected by any of the analyses described below) in comparison with the abundance of the functional groups in the whole set (either proteomic or transcriptomic), as given in Table 2. If the p-value was ≤ 0.05, the test was considered to be significant.

3 Results

3.1 Functional assignment of genes and proteins

The *S. coelicolor* functional genome database contains genomic information about 7825 genes that are assigned into three functional categories that have different levels of specificity. For our purposes, the most appropriate level was the second level, which classifies genes into 27 functional groups (Table 2). Table 2 lists the number of genes that are assigned to 27 functional classes for the microarray experiment and for the proteins that are both Sypro Ruby-stained and radiolabeled.

3.2 Principal components analysis (PCA) of proteomic and transcriptomic experiments

3.2.1 Comparison of the PC loadings

The correlation between mRNA and the protein abundances can be performed by comparing the values relative to a specific fixed time point that has a biological significance [12], and the following points are represented by the log base 2 ratio between the reference point value and the given point value. Such a point, in our case, represents dormant spores. In dormant spores, many of the proteins and mRNAs are not yet synthesized, and the first time point would thus often be represented by a zero value, which would cause the logarithm to approach infinity. Therefore, we could not make a direct comparison between the mRNA and protein abundances. We could only compare the shapes of the expression profiles that represent the kinetics of the expression of individual genes or proteins. All of the individual gene profiles in all of the three experiments were, therefore, normalized to

have a zero mean and are used in further analysis (see paragraph 2.4.1.). A correlation analysis between mRNA and protein kinetics showed higher correlation (Pearson correlation coefficient ≥ 0.95) only for 27.9% of profiles (Supplementary Table 1), while the overall correlation through the dataset was rather low r=0.05. Using the gene annotations and assigning each gene to a diverse functional group (Table 2), a functional analysis of the highly correlated profiles was performed. The functional analysis indicated that there is no bias among the highly correlated profiles toward a specific functional group.

Therefore, instead of focusing on the correlation of individual kinetic profiles, it is more reliable to focus on extracting the common features of the system that are inherent in the expression time series. We chose principal components analysis, which allows identifying representative patterns of kinetic profiles according to their contributions to the overall variance of the dataset. The PCA was performed individually for the microarray experiment and the two proteomic experiments (the PCA or alternative SVD were also used previously in streptomycete gene expression studies [11,12]). The first principal axis loadings (eigenvectors, PCs) bore 30% of the total variability for the microarray and Sypro Ruby-stained proteomic data and 59% for the radiolabeled proteomic data. The first three principal components represented 62%, 55% and 85% of the data variability of the three experiments (Figure 1).

The first principal component loadings profile for the proteomic fluorescently stained experiment is shown in Figure 2a. Surprisingly, there was a striking similarity between the eigenvectors, found, when the eigenvector order of the proteomic fluorescently labeled experiment (PC (Sypro)) was shifted by one. Thus, a good correlation was found between PC1 (mRNA) and PC2 (Sypro), PC2 (mRNA) and PC3 (Sypro), and PC3 (mRNA) and PC4 (Sypro) (Figure 2). Examining the PC1 profile of the Sypro Ruby-stained proteome, a decline in the first two hours followed by a constant level is evident. We can, therefore, speculate that the PC1 (Sypro) is associated with the consumption of those proteins that were stored in the spores. In this case, such a phenomenon cannot be observed, neither on the mRNA nor on the radiolabeled protein levels. We would observe similarity among the eigenvectors only if the first PC (Sypro) is ignored; in other words, if the order of the PCs (Sypro) were shifted by one, then the result is the phenomenon that we indeed observe. Proteins with profiles that are correlated with the 1st PC (Sypro) loading were distributed among the functional groups in the same way as the proteins of the whole

set (data not shown). No bias toward a specific functional group was observed. Therefore, in further PCA analysis, we compared the transcriptomic and Sypro Ruby-stained proteome using this "shifted" order for the Sypro Ruby-stained proteome.

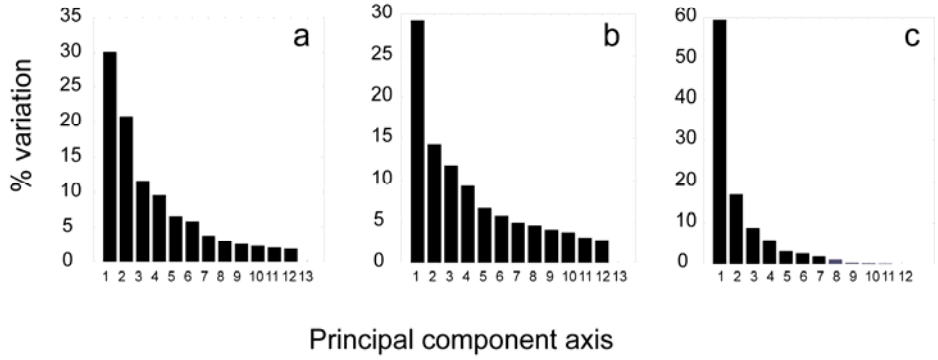


Figure 1. Percentage of variation (eigenvalues) accounted for by each of the 13 principal components. a – transcriptome, b – Sypro Ruby-stained proteome, c – radiolabeled proteome.

A comparison of the first principal component loadings for the three experiments is shown in Figure 2. Figure 2 shows good agreement between the profiles of the eigenvectors of fluorescently labeled proteomic and microarray experiments, documenting common trends in the gene/protein expression kinetics. The radiolabeled proteomic experiment followed the common trend only on the first eigenvector, and higher eigenvectors differed substantially. This difference can be explained by the different nature of the observed data, as the microarrays and Sypro Ruby-stained proteins represent protein or mRNA accumulation, while the pulse radiolabeled proteins represent the rate of protein synthesis. Because the accumulated protein expression level is the result of protein synthesis and degradation, while radiolabeling quantifies only the synthesis, their kinetics should, in principle, differ. The eigenvector profiles confirm that the accumulation kinetics (the result of balance between the synthesis and degradation), which is represented by the first eigenvectors, had a common trend, while the proteosynthesis rate followed different kinetics.

3.3 Gene expression profiles correlated with principal components loadings

Loadings or eigenvectors computed using singular value decomposition, a procedure similar to principal component analysis, that were computed from gene expression profiles, were shown to bear principal information about the kinetics of the processes associated with their shape [11,13,14,15,16]. They allow for the extraction of specific factors from the principal patterns of gene expression within a data matrix through comparisons with independent biological data and are the functional assignments of genes and their products. An analysis of the metabolic networks suggests the existence of different levels of cell control [24] (e.g., most of the metabolic flux in *E. coli* is controlled by only a small number of processes [24]). Such processes form the biochemical backbone for the physiological development; these are the processes that are associated with the programmed development of the cell. On this basic level, other regulatory processes controlling specific metabolic and regulatory activities at specific moments are superimposed. The whole scheme can be depicted as a hierarchy of processes at different levels of specificity. The final kinetics of gene expression is thus determined by the weighted superposition of all of these contributions. If these processes are uncorrelated, then the PCA can deconvolute the information in the profiles of the whole transcriptome or proteome and identify the principal kinetic shapes. Correlation of these shapes, i.e. PC loadings, with individual gene expression profiles and functional annotations of the genes of correlated profiles can reveal the hierarchy of the processes that control the developmental phase under analysis, with the first principal components bearing the most important physiological processes. Deconvolution technically means sorting the PC loadings and their profiles according to their information content level, which is given by their eigenvalues. The first few loadings (usually 3), thus bear most of the information about the kinetics of the underlying processes. The expression profiles of the genes that are correlated with the first principal component loadings thus represent the processes that have the highest importance for the developmental program that they record. Identifying the genes and their functional assignments, whose profiles are correlated with the first few PC loadings, can identify the metabolic and regulatory pathways that are fundamental for the studied developmental processes. In the following section, we focus on the correlation

analysis of gene expression profiles with the first three principal components, i.e., PC1-PC3.

Because the numbers of proteins in the individual functional groups were rather low, in the analysis of the correlation of expression profiles with PC loadings (given in the next paragraphs), we analyzed only the transcriptomic time series.

3.3.1 PC1

The first principal component shows after drop down in the first time point a continuous increase in the gene product accumulation for all three types of data (Figure 2b). PC1 for mRNA (PC2 for Sypro Ruby-stained proteins) increased its accumulation throughout the whole germination period. PC1 for radiolabeled proteins kept increasing almost exponentially. Because all of the expression profiles were normalized to have the same mean, we cannot obtain an absolute value for the expression of individual genes that is correlated with this profile. However, we can presume that the gene products that had the PC1 shape were either already accumulated in the dormant spores or were synthesized immediately after the initiation of germination, and their accumulation kept growing during the 5.5 hours of germination. The Pearson correlation between the PC1 and the gene expression profiles identified 1403 genes (26% of the total) that were significantly correlated ($p<0.05$). A statistical test that compared the abundance of the gene functional groups in the correlated set and the whole set of potentially expressed genes showed that the PC1-correlated set contained over-represented genes for the functional groups “Cell division” (the relative abundance in this set was 2.4x higher than in the full set), “Macromolecule synthesis, modification” (1.6x), “Metabolism of small molecules” (1.73x), “Energy metabolism, carbon” (1.5x), “Ribosome constituents” (2.2x), and “Protein kinases” (2.07x).

The enrichment in the functional groups “Cell division”, “Macromolecule synthesis, modification” and “Ribosome constituents” suggests that the processes represented by PC1 are characterized by the initiation of the spore basal metabolism after breaking the dormancy (as is the translation machinery) and launching the active management with available energy sources (the over-represented “Energy metabolism” group). The list of PC1-correlated genes is given in supplementary Table 2.

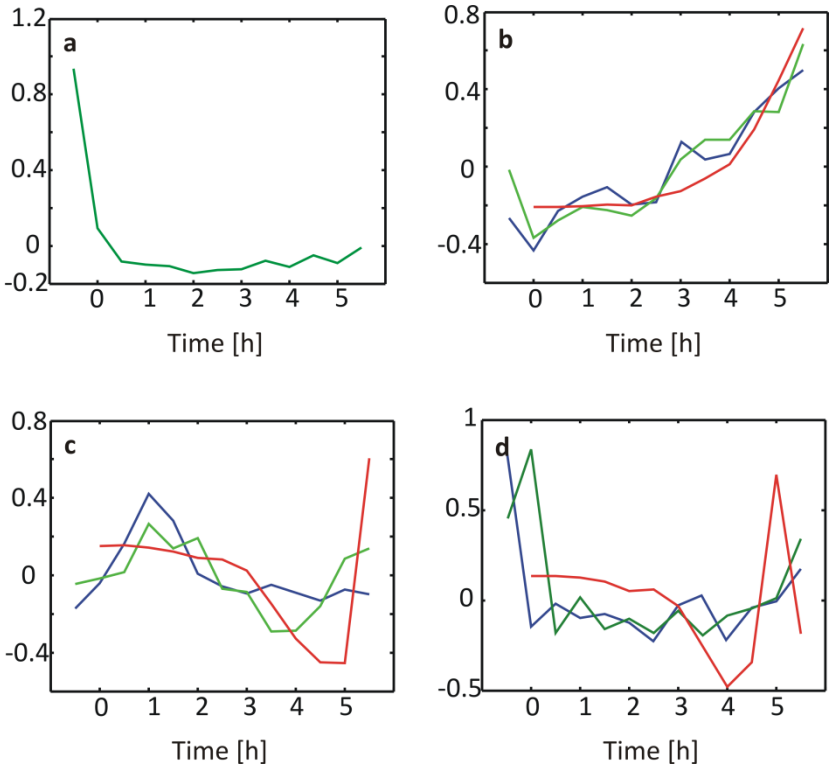


Figure 2. Profiles of the first PC loadings for the transcriptomic experiment and the two proteomic experiments. a) PC1 of the Sypro Ruby-stained proteomic experiment. b) blue – PC1 mRNA, green - PC2 Sypro, red – PC1 radiolabeling. c) blue – PC2 mRNA, green - PC3 Sypro, red –PC2 radiolabeling. d) blue – PC3 mRNA, green - PC4 Sypro, red – PC3 radiolabeling. The first time point in the radiolabeled profile is missing because this point represented dormant spores that could not be radiolabeled.

3.3.2 PC2

The second principal component loading profile was characterized by a peak at approximately time point 4 (1 h) (Figure 2c). Altogether, 791 gene expression profiles were correlated with PC2. Over-represented in this group were the genes that were assigned to groups that are associated with the synthesis of macromolecules, which are either associated with their modification ("Macromolecule synthesis, modification", which was 1.7 times more present in the PC2-correlated set than in the full set of genes) or with providing building material for protein biosynthesis (the functional group "Amino acid biosynthesis" (2x)). Other regulatory proteins that

were not associated with the main regulatory groups (i.e., two component systems and transcription), which are mostly annotated as DNA binding proteins (functional group “Regulation/Other”), were 1.8x over-represented. Additionally, genes of the group “Transport/binding” were significantly over-represented, as were genes of the group “Regulation/Defined families”. Genes that belonged to the “Regulation” group were mostly transcriptional regulators of the R families (such as LysR, TetR, MerR or MarR). Not surprisingly, the genes of the groups “Differentiation/sporulation”, “Cell division”, and “Laterally acquired elements” that are associated with the processes different than germination, were totally absent.

Unlike the genes that correlated with PC1 (which represent basal metabolism), the over-represented groups associated with PC2 (a peak at 1 h) are those that reflect the actual developmental state of the cell and direct the further growth of the cell. In this group, we found an over-represented diverse spectrum of genes that have regulatory functions from the groups “Regulation/Defined families”, “Regulation/Others” and “Transport/binding”. This result indicates that, by means of the regulation factors whose expression peaks are at 1 h, the cells can detect signals from and respond to the environmental conditions. This finding is in agreement with our previous proteomic analysis [3], in which, at the time point preceding the peak at 1 h, we detected the synthesis of most of the regulatory and transport/binding proteins in the germinating spores. The subsequent decrease after the peak at 1 h in the PC2 profile is most likely a result of the synthesizing, transport and regulatory functions of the protein members of the PC2 groups that lead to the controlled expression of the other components that build the growing cell. The list of PC2-correlated genes is given in supplementary Table 3.

3.3.3 PC3

The third principal component loading profile was characterized by an initial maximum at the beginning of the germination and an increase at the end of the measured period (5.5 hours, Figure 2d). A total of 342 genes were found to follow this profile. An analysis of the over-represented functional groups showed more abundant genes in only two groups: “Amino acid biosynthesis” (2.54x) and “Regulation/Protein kinases” (4x). All of the other groups had the same relative presence of the genes as the relative presence that was found in the whole set. The

list of PC3-correlated genes is given in supplementary Table 4. The “Amino acid biosynthesis” group was also enriched among the genes that correlated with PC2, and the “Regulation/Protein kinases” group was over-represented in the genes that correlated with PC1. In contrast to PC1 (increasing), the character of the PC3 kinetic trend (an abrupt decrease in the first hour) suggests that a rapid switch occurred within the first hour and was mainly in the expression of the genes from the two over-represented groups. The switch in the usage of the amino acid biosynthesis group in the first hour can be explained as a reaction of the metabolism to a sudden supply of amino acids from the AM medium that was used for cultivation, when the cell was adjusting its metabolism to the current environmental conditions. Metabolic changes might also be associated with the requirement of a different specific set of protein kinases compared with the set that is needed just after the germination activation.

Principal component analysis of the gene expression data showed that the most important aspects are the processes that are associated with the first three principal components. Unlike in our previous work [11], where we found a strong association between the fifth PC and antibiotic production, the higher principal components could not be associated with any developmental process (data not shown). The number of genes that were associated with the principal components decreased rapidly with the decreasing eigenvalues of the corresponding principal components. A total of 1403 gene expression profiles were correlated with PC1, 791 were correlated with PC2 and 342 were correlated with PC3. The experiment was also designed to cover the biological variability that occurred when each repeat from each time point measurement was collected from different biological replicates (see the Methods section). Although it was not statistically proven, the biological variability was the main source of the experimental error, and the lowest PCs were apparently associated with the experimental noise. Our analysis proposed that the PC1-associated processes are connected with launching the basal metabolism (a more detailed analysis of the basic metabolic processes supporting this statement is given in the following paragraph). Comparison of PC1 profile with the course of DNA synthesis (Supplementary Figure 2), where approximately at 2.5 hours of growths DNA replication starts, shows moderate correlation with PC1 for microarray and proteomic Sypro experiments, indicating association between PC1 correlated genes/proteins and first DNA replication. Making a definite statement about the

correlation between PC1 associated genes and DNA synthesis is complicated by rather high variance of both proteomic and transcriptomic experiments reflected in the PC1 profile. Although it cannot be undoubtedly confirmed such observation has to be mentioned.

The PC2-correlated processes represent the response of the cell to the actual environmental and/or inner conditions through corresponding regulatory pathways. PC3 reflected processes that are suppressed after germination initiation as a reaction to the medium composition detected by the cell. Genes that are associated with these processes were identified and are available in the supplementary materials.

3.4 PCA and major functional groups

In the preceding paragraphs, genes in the major functional groups were examined for their correlations with the first principal eigenvectors (PC1). In the next section, we focus on genes that are essential for re-activating metabolism after breaking dormancy, i.e., the genes of energy metabolism, nucleic acids and protein synthesis, and their association with the principal kinetic shapes represented by first principal eigenvectors. In addition, the genes of the stress response were examined because germination can be considered a reaction to the stress that is associated with the rehydration of the originally dry spores.

3.4.1 Energy metabolism

We use the specific gene annotation and a pathway mapping tool given in the KEGG database (Kyoto Encyclopedia of Genes and Genomes, http://www.genome.jp/kegg-bin/show_organism?menu_type=pathway_maps&org=sco) to investigated mRNA expression profiles involved in the primary energy metabolism. The most of the energy pathways-associated genes are significantly correlated with PC1 (Figure 3), including genes comprehended in TCA cycle, pentose phosphate pathway and glycolysis. Figure 3 shows mapping of PC1 correlatad gene expression profiles on to the KEGG pathway map.

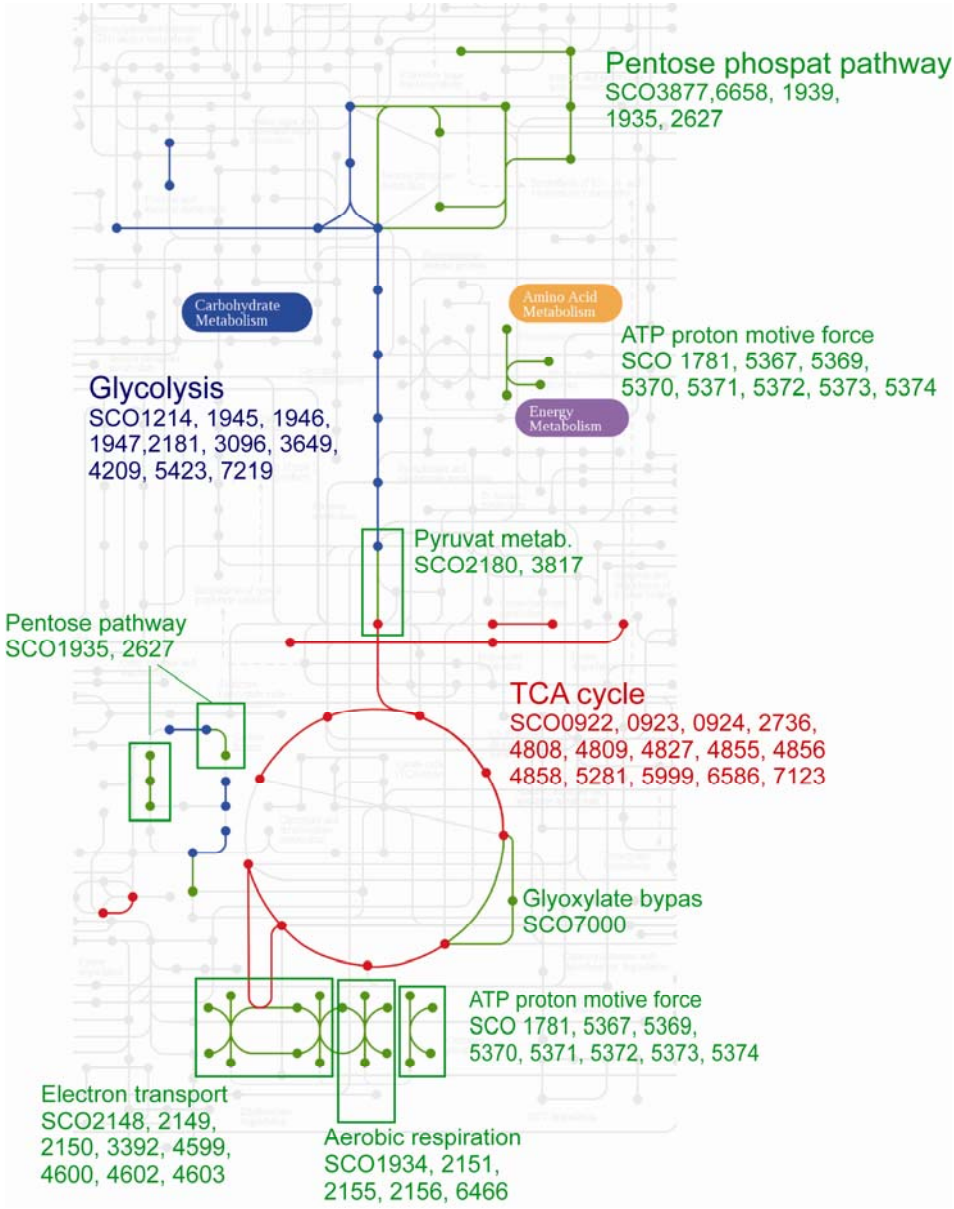


Figure 3. A visualization of genes significantly correlated with PC1, which are involved in primary energy metabolism. The basis of the illustration was made in KEGG mapping tool.

For the genes annotated in KEGG, the significant correlation with PC1 was found for 38% of TCA cycle genes, including the subunits of succinyl CoA synthase (SCO4809,6586), citrate synthase (SCO2736), malate dehydrogenase (SCO4824) and acetyl transferase (SCO7123), for 22% of glycolysis genes and 25% of genes from pentose phosphate pathway.

A significant correlation with PC1 was also found for mRNAs that encode for the pathway that leads from the glyceraldehydes 3-phosphate to pyruvate (*gap*

(SCO1947), *pgk* (SCO1946), and *eno* (SCO3096)) and for the oxidative phosphorylation genes *cox1* (SCO2155, 2156) and *qcrB* (SCO2148) and ATP synthase (SCO5374). For none of the genes from TCA cycle, pentose phosphate pathway and glycolysis a correlation with PC2 was found.

3.4.2 Stress response

The systematic annotation of stress response genes is not available; thus, we extracted the names of the stress response genes from the two most relevant resources, i.e., GenBank and StreptoDB (http://strepdb.streptomyces.org.uk). Altogether, 46 genes that represent heat, cold and starvation shock were analyzed. Approximately half of them were correlated with PC1. Heat shock proteins did not show any correlation with PC1. In contrast, 6 of the 8 cold shock proteins that were found were positively significantly correlated with PC1 (SCO4684, 527, 4505, 5921, 3748, and 3731). The general stress protein 50S ribosomal protein L25 *ctc* (SCO3124) had the highest correlation with PC1 (r=0.89) among all of the stress proteins. Catalase *catA/C* (SCO0379, 0560) showed a negative correlation with PC1. Starvation genes (*pstB* (SCO4139), *pstS* (SCO4142), *regX3* (SCO4230), a phosphate transporter (SCO4228), alkaline phosphatase (SCO2286, 1906, 3790, 5140) were mostly negatively correlated or not correlated with PC1. Only *regX3* (SCO4230) and dehydrogenase (SCO2490), a general stress protein, were found to be significantly correlated with PC2. Other stress proteins did not show any significant correlation with PC2.

3.4.3 Nucleic acids and protein synthesis

Proteosynthetic genes were highly correlated with PC1. Of 53 proteosynthetic genes (groups "Amino acyl tRNA synthase tRNA modification" and "Proteins - translation and modification"), 80% were significantly correlated; the correlated genes primarily comprised tRNA synthase genes and the genes for elongation factors Ts (SCO5625), Tu (SCO4662), P (SCO1491) and G (SCO4661) and the translation initiation factors IF-1/2 (SCO 4725, 5706). In contrast, RNA synthesis and DNA replication genes were correlated only moderately (30% of 113 were significantly positively correlated). Among the positively correlated genes, the subunits of DNA

polymerase III (SCO1827, 2003, 6084, 3541) or helicase (SCO 2952, 1167) were found.

The correlation with PC2 was always lower than that of the PC1 genes and was highest for the genes that were involved in the "DNA replication repair" group (26% of 80).

The analysis of the association of specific functional and metabolic groups with the first two eigenvectors showed that PC2 did not correlate with almost any of the selected groups and genes, while PC1 was correlated with specific functional groups. As PC1 represents principal kinetic shape controlling germination, genes correlated with PC1 represent principal pathways controlling germination. The principal expression profile (Figure 2b) shows that the genes having this profile are relatively highly expressed in dormant spores, their expression drops just after germination initiation and after that, their relative expression continuously increases until the end of measured period. Expression profiles having this shape belonged mostly to the genes of energy metabolism that are involved in basic metabolic processes, such as the TCA cycle and its associated pathways Genes that were highly correlated with PC1 were involved in nucleic acids and protein synthesis, including the main translation factors. In agreement with the results of the global analysis of the association of general functional groups with eigenvectors (paragraph 3.3.), the basic processes of germination, with the kinetics represented by PC1, involve biochemical and regulatory pathways that are indispensable in accelerating the primary energy metabolism to an increased level and are required for the cell to become competent to develop into vegetative forms. These data also indicate that, as a response to the increased demand of the new proteome constitution and, therefore, the capacity of a proteosynthetic apparatus, the sole re-activation of aggregated proteosynthetic components is insufficient and must be accompanied by the *de novo* synthesis of its members.

From the group of stress responses, the genes associated with cold shock were highly correlated with PC1. This result confirms the hypothesis suggested by Strakova et al. [3], i.e., that germination involves gene expression processes that resemble cell responses to stress conditions.

3.5 Absolute gene/protein expression levels

It was shown above that the genes associated with PC1 bore principal expression profiles associated with genes driving germination. The remaining question was, whether the PC1 associated genes had different levels of expression in comparison with the whole mRNA dataset.

The samples on the chips in a transcriptomic experiment were labeled and organized such that the samples were in one channel (the alpha channel, labeled with one fluorescent dye) and the standard (a mixture of samples from all time points, see Methods) were in the other channel (the beta channel, with a second fluorescent dye). The time series in all of the above analyses used the ratio between the alpha and beta channels. This arrangement decreases the measurement variance but does not allow for a direct comparison of the absolute expression levels among the different genes because the standard hybridizes differently with different mRNA probes that are immobilized on the chip. Because the mRNA sample repeats were randomly distributed among the different arrays and equal amounts of mRNA were always loaded on the chip, we were able to use the alpha channel alone and compare the absolute expression levels of the individual genes. The absence of the standard increases the variance of the averaged expression values of individual mRNA levels but should not influence the overall trend; in other words, the expression profiles obtained both from the alpha channel only and from the normalized data should be correlated. An analysis of the correlation coefficient distribution between the alpha channel and normalized ratios showed that 75.6% of the genes were significantly correlated ($p<=0.05$), with a maximum at $r=0.87$. We also found that the lowest correlation was associated with profiles that had an overall low expression level in the profile.

For the individual genes from the alpha channel (representing the absolute gene expression level) the log-base2 distribution of the medians of the expression time series is shown in Figure 4. This distribution, even in the logarithmic scale, was heavily tailed toward highly expressed genes. When compared with the number of identified proteins (the thick bars in Figure 4) that correspond to the genes in a given quartiles of the distribution, it is apparent that most identified proteins were within the fourth quartile of the distribution, especially in the most highly expressed 5% of

the genes. The relationship between the expression level of mRNA and the number of corresponding proteins that are expressed is apparent.

When examining the functional characteristics of the genes in the most highly expressed 5% percent, the genes were significantly enriched in the groups "Regulation/RNApolymerase core enzymes binding protein" (2.4 x), "Ribosomal constituents" (which correspond to 80% of all of the ribosomal genes predicted in the *S. coelicolor* genome), and "Chaperones" (60% of all of the genes predicted in the genome). The possible role of chaperones in germination was discussed in the work of Bobek et al. [8]. There were also significantly more genes that were involved in the translational machinery (the elongation factors, RNA polymerase subunits), the energy metabolism and, strikingly, the cold shock genes.

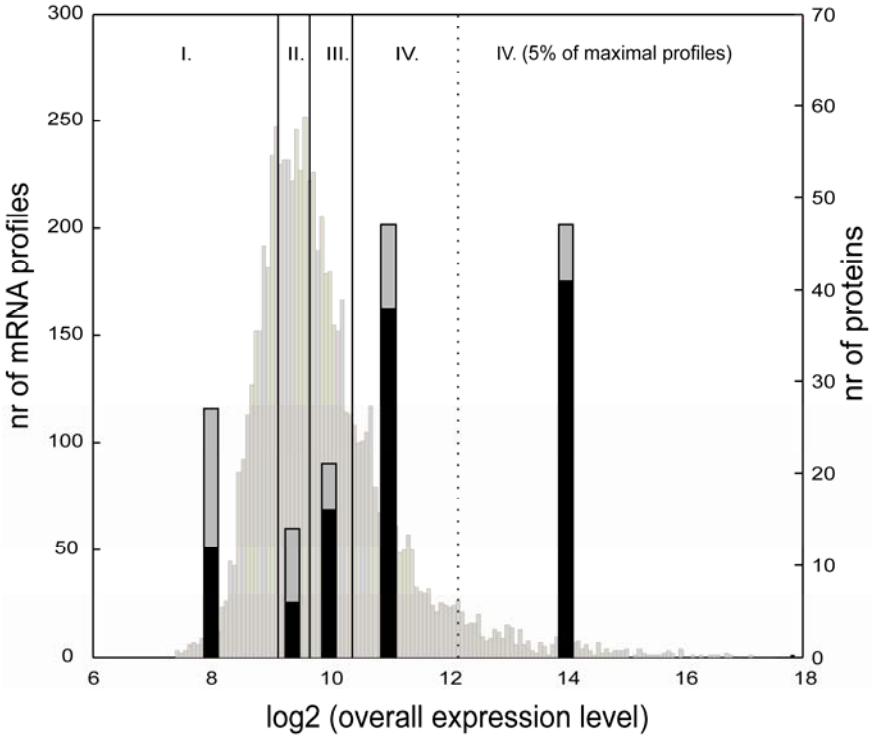


Figure 4. Distribution of the medians of the gene expression profiles, given in absolute units, as measured in the microarray alpha channel. The distribution is divided into 4 quartiles (roman numbers) and the last 5% of the most highly expressed genes. Vertical bars represent the number of proteins of Sypro Ruby-stained (red bars) or radiolabeled (black bars), in the proteome that corresponds to the genes in the given quartile.

Comparing the analysis of the association of genes with the first principal components (paragraph 3.3.), a similarity between the highly expressed group and the group of genes that are correlated with the PC1can be observed. The similarity implies that the principal regulatory groups, which are necessary for germination control and progression, have not only the kinetic profile defined by the PC1 shape but also belong to the overall highly expressed group. The presence of cold shock genes in both groups is interesting. As mentioned in [25] bacterial cold shock proteins sequences are conserved among species, however their role can differ in various organisms and furthermore not all of them are induced by cold shock. As the genome annotation is frequently made by homology search, the annotation of the genes as cold shock genes may be caused by the way the genes were annotated. We have checked all the cold shock genes mentioned above for their appearance in the literature, but we didn't find any particular reference dealing with their function in *Streptomyces*. It is apparent that these genes are essential for the progression of germination, but their actual role remains a puzzle, and it will require further work to determine their function in germination.

Detailed inspection focusing on individual genes shows that the group of most expressed genes comprised important transcription regulators: sigma factors, anti-sigma factors and anti-anti sigma factors. Interestingly, we detected a very large expression of the gene for the alternative principal sigma factor HrdD (SCO3202), whose function has not yet been revealed in *S. coelicolor*. The group also contained both genes for SigH anti-sigma factor Prs (SCO5244) and the gene for its partner switch, an anti-sigma factor antagonist BldG (SCO3549), which were recently shown to interact directly and participate in switching-like activation/deactivation of the sigma factor SigH [26]. Similarly, the regulation based on a switch-like mechanism was proposed for anti-anti-sigma factor ArsI (SCO3067), the SigI anti-sigma factor antagonist [27]. In addition to the gene for an extracytoplasmic function (ECF) sigma factor SigE (SCO3356), which regulates genes that are involved in cell wall biosynthesis [28] among the maximally expressed genes of regulators, *sigD* (SCO4769) and genes predicted to encode sigma factors SCO0038 and SCO4908 (which products have still unspecified function) were found. Finally, the detected members of this enhanced regulatory group were the genes for co-expressed partners, sigma factor SigR (SCO5216) and its anti-sigma factor RsrA (SCO5217), which together create a control system that is sensitive to changes in the intracellular redox

balance [29]. Interestingly, in the enriched functional group "Biosynthesis of cofactors and carriers" (2x), we found several genes that were previously reported as SigR target genes [30,31], such as thioredoxin enzymes (SCO0885, SCO1084, SCO3889), which assist in reducing disulfide bonds that are unnatural for the intracellular environment. Additionally, genes that were involved in the translation machinery were found in the highest expressed gene group, including the genes for elongation factors P, G, Tu, Ts (SCO1491, 4661, 4662,5625) and initiation factors IF-1 and 2 (SCO4725, 5706). As mentioned above, among the highly expressed genes, the cold shock genes (SCO527, 3731, 4295, 4505, 4684, 5921) were found.

3.6 Analysis of coding gene utilizations

Further, the changes in number of highly expressed genes during germination within individual functional groups were investigated.

We used data from the alpha channel, as described in the previous paragraph. To select genes that were highly expressed and to eliminate the influence of the high variance of the low-expressed genes on the analysis, we chose the threshold level of the third quartile of the distribution calculated from alpha channel signals. All of the mRNAs that had a signal higher than this threshold were selected for further analysis. Then, for each time point, all of the genes with an expression value above this threshold were identified and sorted into 27 functional groups (as defined in Table 2). The numbers of highly expressed genes assigned into each individual functional group were counted at each of the 13 time points, generating a time series of the number of highly expressed genes in the individual functional groups (Figure 5, the blue curve). To validate whether the threshold value (defining high expression genes) could influence the profile shapes of the numbers of expressed genes, the same process was repeated with a threshold values that were equal to the first and second quartile (data not shown). The results showed that the selection of the threshold value has no effect on the shapes of the profiles given in Figure 5, and therefore, we used the original threshold value of the third quartile.

The resulting profiles are shown in Figure 5. Figure 5 shows that the overall number of expressed genes during the course of germination exhibited two peaks (Figure 5, caption "Total"); the first occurred in T0, and second occurred after approximately 2.5 hours of growth (time points 6-7). After the second peak, the

expression stabilized at a level of approximately 1200 highly expressed genes. It is striking that there is a relatively high number of mRNAs that are found in dormant spores (ca. 1600). A previous study of *S. granaticolor* in experiments using the transcription inhibitor rifamycin revealed that dormant spores preserve pre-existing mRNAs, which are expressed at the beginning of germination [5]. In our previous work on the proteomic dataset [3], we found that several newly synthesized proteins appear just minutes after germination initiation. The finding that a relatively large number of mRNAs exist already in spores could suggest that these proteins were synthesized from this stock.

To compare the trends in individual functional groups with the overall trend, the individual profiles were divided by the general pattern and were multiplied by 100 (red curves in the graphs of Figure 5). Such curves show a deviation from the general pattern (caption "Total" in Figure 5) for the given functional group. The group of "Regulation/Protein kinases" was excluded because it contained only a few mRNAs (0-3). Additionally, the groups of "Cell division" and "Differentiation/sporulation" contained a small number of mRNAs (5 on average), but they were retained for comparison.

The time series of the abundance of the genes of different functional groups can be dissected into 5 principal patterns according to the development of the relative number of genes (red curve in Figure 5) that were expressed during germination. The first and largest group of functional assignments copied the profile of the general pattern (blue curve and flat red curve in first column of graphs in Figure 5). It can be expected that the genes in the "Unknown function" and "Not classified" groups will follow the general pattern because they contain an uncharacterized mixture of genes. Aside from the above-mentioned functional groups, this group contained mainly regulatory genes and genes involved in the synthesis of macromolecules and amino acids.

The second largest group showed an increase in the relative number of expressed genes over time (red curves in second column of graphs in Figure 5), including the genes of "Energy metabolism", "Central intermediary metabolism" and "Fatty acid biosynthesis".

The third group was formed by the genes that were expressed always in the same numbers, regardless of the development phase (blue curves in third column of graphs in Figure 5). Not surprisingly, this group was formed by "Ribosome constituents",

which are expected to be constitutively expressed, and “Chaperones”. This group also contained the genes of “Nucleotide biosynthesis” and “Macromolecular synthesis”, which similarly followed a constant trend.

Groups of “Macromolecular degradation”, “Laterally acquired elements” and “Transport/binding” followed the general trend but had a much higher emphasis.

The genes of “Secondary metabolism” were surprisingly expressed in non-negligible numbers (approximately 30 (12.5%) of all secondary metabolism genes), and its numbers declined over time and had a maximum in the dormant spores. These transcripts usually originate from the sporulation stage, in which the antibiotics are produced, and they are subsequently degraded. The presence of several members of antibiotic gene clusters might also suggest that their enzymatic activity is required in germination. However, none of those proteins were detected here. A detailed inspection of the individual genes did not reveal any specific and/or continuous gene clusters for the synthesis of secondary metabolites. Because the group of “Secondary metabolism”, as defined by the Sanger Institute, also includes a number of other genes that are not directly associated with biosynthetic clusters (such as lipoproteins), an over-representation of this group is most likely not associated with secondary metabolite production but is instead associated with those genes that are not directly involved in the synthesis of secondary metabolites.

Similar pattern could also be observed for the genes of the group “Degradation of small molecules”.

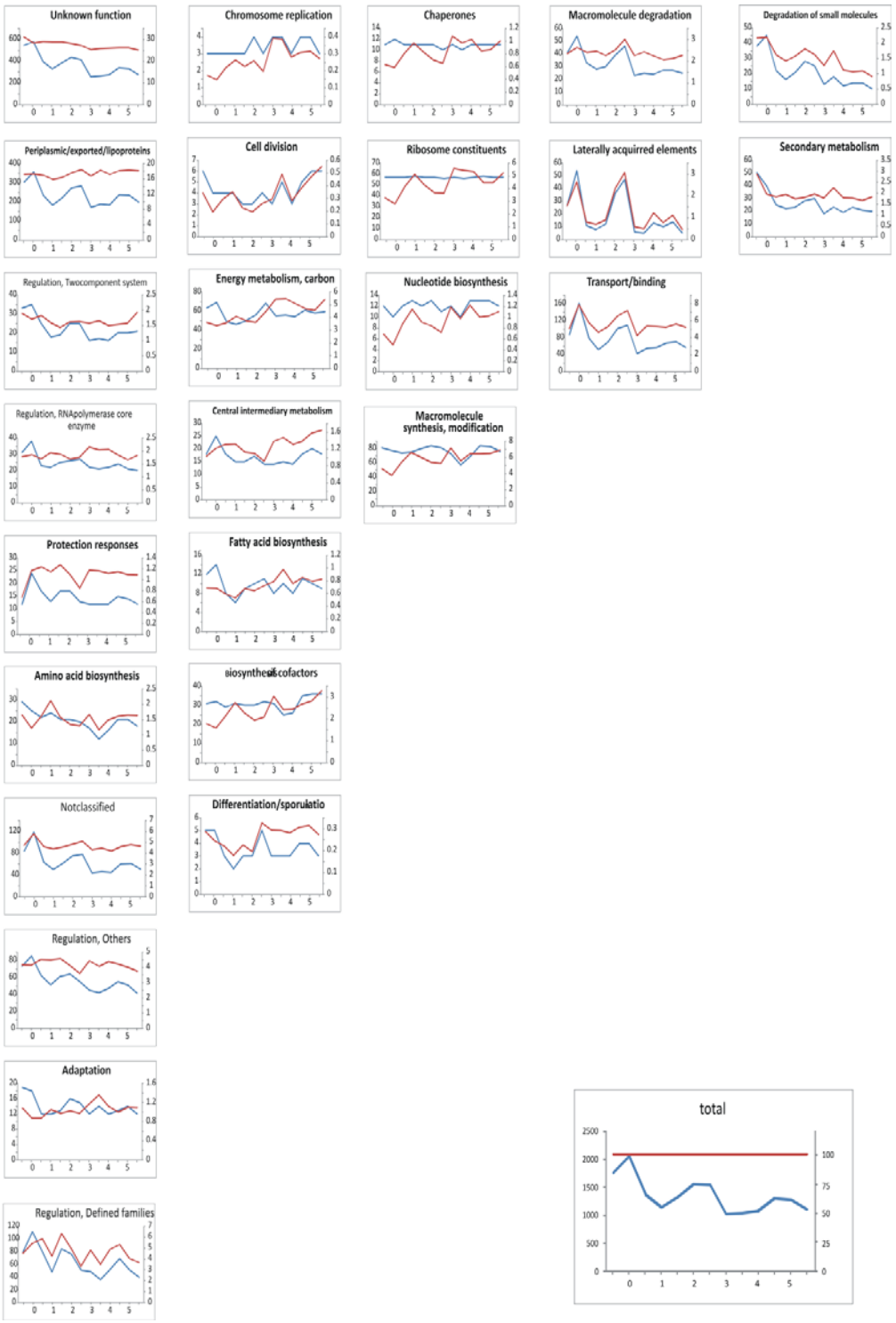


Figure 5. The number of expressed genes of different functional groups over the course of germination. Blue curve (left vertical axis) - absolute numbers of expressed genes at a given time point, red curve (right vertical axis) - number of expressed genes relative to all of the expressed genes, in terms of the percent. Horizontal axis – time [hours]. Individual functional groups are ordered in columns according to the similarity of the profiles.

4 Discussion

A correlation analysis between proteomic and transcriptomic data showed a rather low correlation between the mRNA and the accumulated protein expression profiles. Of the 247 genes/proteins that were investigated, 27.9% were highly correlated (a correlation interval of ≥0.95). This finding is consistent with other analyses on *Streptomyces* species, which also found a correlation for approximately one-third of the genes that were expressed during the stationary phase [10,11,12]. Several similar studies, which were mostly performed in yeast [32,33,34,35], have reported varying, but still rather low, correlations between mRNA and protein abundances, which range from 0.21 to 0.74 (Pearson correlation). The variability between datasets mostly goes to the account of the translation and posttranslational processes and also the errors in the measurements, which are quite high.

Whenever a comparison between proteomic and transcriptomic data is performed, two sets that substantially differ in size are compared. While proteomic experiments, which are made either by 2D electrophoresis or mass spectrometry, can quantify usually hundreds of proteins, microarrays provide information about thousand of genes. The sets differ in size by one order of magnitude, which leads to a comparison of biased selection of only highly abundant proteins with unbiased microarray expression data. We must admit that, except for the analyzed intersect, we do not and will not know what is the "true" correlation (across the whole proteome and transcriptome), until the proteomics will be able to quantify a more representative part of the real proteome. Therefore, instead of focusing on the correlation of individual expression profiles, it is more reliable to focus on extracting the common features of the system that are inherent in the gene expression time series. One of these methods is the principal components analysis, which, when used to analyze a set of gene expression profiles, allows us to identify patterns of gene expression that contribute to the pathways controlling the observed process. Comparing these patterns for different experiments (proteomic or microarray) could say how much these methods identify common features of the system and could show how much they give similar results. The analyses made so far on *Streptomyces* [11,12] show very good agreement between proteomic and transcriptomic temporal data at the level of the first eigenvectors, which confirms that the fundamental processes are controlled in a coordinated fashion on both the transcriptomic and

proteomic levels. The pulse-labeled data differed substantially, which is most likely caused by the different nature of the measured values – the accumulation of the mRNA and protein fluorescence staining vs. the expression rate for the pulse radiolabeling. By correlating individual gene expression profiles with the first eigenvectors, we were able to identify the metabolic and regulatory pathways that control the fundamental processes during the germination of *S. coelicolor*.

The analysis of highly expressed genes showed a correlation between the expression levels of the mRNAs and the corresponding proteins; genes that are highly expressed on the mRNA level are also highly expressed on the protein level. Functional analysis of most of the highly expressed genes showed that the highly expressed genes are also those that are correlated with PC1. This comparison shows that the principal regulatory groups that are necessary for the germination control and progression are, overall, highly expressed and have a kinetic profile that is defined by the shape of the first principal component.

Respecting the limitations of the proteomics approach, we utilized a more detailed analysis that addresses individual genes and functional groups, and we focused on mRNA data and interpreted the gene expression in the sense of genome utilization over the course of the experiment. The analysis of the number of genes of individual functional groups that were expressed during the course of germination showed that a relatively high number of mRNAs existed already in the dormant spores. The mRNA synthesis peaked at the first time point of the germination and declined until the end of the observed period, with a small local maximum at 2.5 hours. This result, which is in agreement with previous proteomic analysis [3], suggests that dormant spores already contain most of the genetic material that is necessary for the spore germination initiation that was stored, most likely in aggregates, that stabilize both the mRNA and proteins and which, after rehydration, can readily initiate the growth of the spores. The peak of the macromolecule synthesis that was found at the first 30 min after initiation supports this statement.

In interpreting both the transcriptomics and proteomics data, we could not go below a certain level of generality given by the inherent information content that is different for both of the data sources. However, analyzing large-scale gene expression data using statistical methods such as PCA can give us insights into how biochemical and regulatory processes work in the cell on the systems level.

5 References

1. Ranade N, Vining LC (1993) Accumulation of intracellular carbon reserves in relation to chloramphenicol biosynthesis by Streptomyces venezuelae. Can J Microbiol 39: 377-383.
2. Ghorbel S, Smirnov A, Chouayekh H, Sperandio B, Esnault C, et al. (2006) Regulation of ppk expression and in vivo function of Ppk in Streptomyces lividans TK24. J Bacteriol 188: 6269-6276.
3. Strakova E, Bobek J, Zikova A, Rehulka P, Benada O, et al. (2012) Systems Insight into the Spore Germination of Streptomyces coelicolor. Journal of Proteome Research.
4. Mikulik K, Janda I, Maskova H, Stastna J, Jiranova A (1977) Macromolecular synthesis accompanying the transition from spores to vegetative forms of Streptomyces granaticolor. Folia Microbiol (Praha) 22: 252-261.
5. Mikulik K, Bobek J, Bezouskova S, Benada O, Kofronova O (2002) Expression of proteins and protein kinase activity during germination of aerial spores of Streptomyces granaticolor. Biochem Biophys Res Commun 299: 335-342.
6. Cowan AE, Koppel DE, Setlow B, Setlow P (2003) A soluble protein is immobile in dormant spores of Bacillus subtilis but is mobile in germinated spores: implications for spore dormancy. Proc Natl Acad Sci U S A 100: 4209-4214.
7. Haiser HJ, Yousef MR, Elliot MA (2009) Cell wall hydrolases affect germination, vegetative growth, and sporulation in Streptomyces coelicolor. J Bacteriol 191: 6501-6512.
8. Bobek J, Halada P, Angelis J, Vohradsky J, Mikulik K (2004) Activation and expression of proteins during synchronous germination of aerial spores of Streptomyces granaticolor. Proteomics 4: 3864-3880.
9. Mikulik K, Zhulanova E, Kratky M, Kofronova O, Benada O (2000) Isolation and characterization of dcw cluster from Streptomyces collinus producing kirromycin. Biochem Biophys Res Commun 268: 282-288.
10. Abreu RD, Penalva LO, Marcotte EM, Vogel C (2009) Global signatures of protein and mRNA expression levels. Molecular Biosystems 5: 1512-1526.
11. Vohradsky J, Branny P, Thompson CJ (2007) Comparative analysis of gene expression on mRNA and protein level during development of Streptomyces cultures by using singular value decomposition. Proteomics 7: 3853-3866.
12. Jayapal KP, Philp RJ, Kok YJ, Yap MG, Sherman DH, et al. (2008) Uncovering genes with divergent mRNA-protein dynamics in Streptomyces coelicolor. PLoS One 3: e2097.
13. Alter O, Brown PO, Botstein D (2000) Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci U S A 97: 10101-10106.
14. Alter O, Brown PO, Botstein D (2003) Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. Proc Natl Acad Sci U S A 100: 3351-3356.
15. Holter NS, Maritan A, Cieplak M, Fedoroff NV, Banavar JR (2001) Dynamic modeling of gene expression data. Proc Natl Acad Sci U S A 98: 1693-1698.
16. Holter NS, Mitra M, Maritan A, Cieplak M, Banavar JR, et al. (2000) Fundamental patterns underlying gene expression profiles: simplicity from complexity. Proc Natl Acad Sci U S A 97: 8409-8414.
17. Omberg L, Golub GH, Alter O (2007) A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies. Proceedings of the National Academy of Sciences of the United States of America 104: 18371-18376.
18. Ramsden JJ, Vohradsky J (1998) Zipf-like behavior in procaryotic protein expression. Physical Review E 58: 7777-7780.

19. Vohradsky J, Ramsden JJ (2001) Genome resource utilization during prokaryotic development. Faseb J 15: 2054-2056.
20. Kieser T, Bibb MJ, Buttner MJ, Chater KF, Hopwood DA (2000) Practical Streptomyces genetics. Norwich, UK.
21. Molloy MP, Brzezinski EE, Hang J, McDowell MT, VanBogelen RA (2003) Overcoming technical variation and biological variation in quantitative proteomics. Proteomics 3: 1912-1919.
22. Krasny L, Tiserova H, Jonak J, Rejman D, Sanderova H (2008) The identity of the transcription +1 position is crucial for changes in gene expression in response to amino acid starvation in Bacillus subtilis. Mol Microbiol 69: 42-54.
23. Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30: 207-210.
24. Almaas E, Kovacs B, Vicsek T, Oltvai ZN, Barabasi AL (2004) Global organization of metabolic fluxes in the bacterium Escherichia coli. Nature 427: 839-843.
25. Kim MJ, Lee YK, Lee HK, Im H (2007) Characterization of cold-shock protein A of Antarctic Streptomyces sp. AA8321. Protein J 26: 51-59.
26. Sevcikova B, Rezuchova B, Homerova D, Kormanec J (2010) The anti-anti-sigma factor BldG is involved in activation of the stress response sigma factor sigma(H) in Streptomyces coelicolor A3(2). J Bacteriol 192: 5674-5681.
27. Homerova D, Sevcikova B, Rezuchova B, Kormanec J (2012) Regulation of an alternative sigma factor sigmaI by a partner switching mechanism with an anti-sigma factor PrsI and an anti-anti-sigma factor ArsI in Streptomyces coelicolor A3(2). Gene 492: 71-80.
28. Paget MS, Chamberlin L, Atrih A, Foster SJ, Buttner MJ (1999) Evidence that the extracytoplasmic function sigma factor sigmaE is required for normal cell wall structure in Streptomyces coelicolor A3(2). J Bacteriol 181: 204-211.
29. Kang JG, Paget MS, Seok YJ, Hahn MY, Bae JB, et al. (1999) RsrA, an anti-sigma factor regulated by redox change. EMBO J 18: 4292-4298.
30. Paget MS, Kang JG, Roe JH, Buttner MJ (1998) sigmaR, an RNA polymerase sigma factor that modulates expression of the thioredoxin system in response to oxidative stress in Streptomyces coelicolor A3(2). EMBO J 17: 5776-5782.
31. Kim MS, Dufour YS, Yoo JS, Cho YB, Park JH, et al. (2012) Conservation of thiol-oxidative stress responses regulated by SigR orthologues in actinomycetes. Mol Microbiol 85: 326-344.
32. Maier T, Guell M, Serrano L (2009) Correlation of mRNA and protein in complex biological samples. Febs Letters 583: 3966-3973.
33. Griffin TJ, Gygi SP, Ideker T, Rist B, Eng J, et al. (2002) Complementary profiling of gene expression at the transcriptome and proteome levels in Saccharomyces cerevisiae. Mol Cell Proteomics 1: 323-333.
34. Gygi SP, Rochon Y, Franza BR, Aebersold R (1999) Correlation between protein and mRNA abundance in yeast. Mol Cell Biol 19: 1720-1730.
35. Schmidt MW, Houseman A, Ivanov AR, Wolf DA (2007) Comparative proteomic and transcriptomic profiling of the fission yeast Schizosaccharomyces pombe. Mol Syst Biol 3: 79.

6 Acknowledgements

We would like to thank Oldrich Benada and Olga Kofronova (Institute of Microbiology, Czech Academy of Sciences) for the electron microscopy images.

7 Tables

Table 1: The number of 2DE gel images and microarrays that correspond to biological replicates analyzed for individual time points of the experiment.

Table 1.

Time [hours]	Number of Sypro Ruby-stained images	Number of radiolabeled images	Number of microarrays
Dormant	4	0	3
0	4	4	3
0.5	4	4	3
1	4	4	2
1.5	3	3	3
2	4	4	3
2.5	3	3	2
3	5	5	3
3.5	5	5	3
4	5	5	3
4.5	5	5	3
5	4	4	3
5.5	4	4	3

Table 2: Functional classification of the genes of the *S. coelicolor* genome according to The Sanger Institute (ftp://ftp.sanger.ac.uk/pub/S_coelicolor/classwise.txt)

Table 2.

	Fluorescent labeling	Radio labeling	mRNA
Unknown function	46	26	1496
Chromosome replication	1	1	8
Chaperones	16	13	14
Protection responses	8	4	54
Transport/binding proteins	30	21	435
Adaptation	4	1	31
Cell division	2	2	14
Differentiation/sporulation	1	1	10
Macromolecule degradation	18	6	152
Macromolecule synthesis, modification	22	15	209
Amino acid biosynthesis	1	0	99
Biosynthesis of cofactors, carriers	6	6	88
Central intermediary metabolism	3	3	78
Degradation of small molecules	4	3	142
Energy metabolism, carbon	25	17	152
Fatty acid biosynthesis	1	0	45
Nucleotide biosynthesis	0	0	28
Secondary metabolism	0	0	163
Periplasmic/exported/lipoproteins	32	13	944
Ribosome constituents	5	4	60
Laterally acquired elements	0	0	76
Regulation/Two component system	3	3	121
Regulation/RNApolymerase core enzyme binding	4	4	68
Regulation/Defined families	3	2	325
Regulation/Protein kinases	0	0	35
Regulation/Others	5	4	175
Not classified (including putative assignments)	11	2	363
SUM	251	151	5385

8 Supplementary materials

Supplementary Table 1 - Genes with highly correlated scores (Pearson correlation >=0.95) of Sypro Ruby-stained proteins (PC2) and mRNA (PC1).

Supplementary Table 2 - Genes that have an expression profile that is correlated with PC1.

Supplementary Table 3 - Genes that have an expression profile that is correlated with PC2.

Supplementary Table 4 - Genes that have an expression profile that is correlated with PC3.

Supplementary Figure 1 - Physiological change of spores during germination.

Supplementary Figure 2 – Comparison of the principal components with DNA synthesis.