J Ind Microbiol Biotechnol (2006) 33: 159–172
DOI 10.1007/s10295-005-0034-7

ORIGINAL PAPER

**Sarika Mehra · Wei Lian · Karthik P. Jayapal
Salim P. Charaniya · David H. Sherman · Wei-Shou Hu**

A framework to analyze multiple time series data: A case study with *Streptomyces coelicolor*

Received: 11 April 2005 / Accepted: 24 May 2005 / Published online: 11 October 2005

Abstract Transcriptional regulation in differentiating microorganisms is highly dynamic involving multiple and interwinding circuits consisted of many regulatory genes. Elucidation of these networks may provide the key to harness the full capacity of many organisms that produce natural products. A powerful tool evolved in the past decade is global transcriptional study of mutants in which one or more key regulatory genes of interest have been deleted. To study regulatory mutants of *Streptomyces coelicolor*, we developed a framework of systematic analysis of gene expression dynamics. Instead of pair-wise comparison of samples in different combinations, genomic DNA was used as a common reference for all samples in microarray assays, thus, enabling direct comparison of gene transcription dynamics across different isogenic mutants. As growth and various differentiation events may unfold at different rates in different mutants, the global transcription profiles of each mutant were first aligned computationally to those of the wild type, with respect to the corresponding growth and differentiation stages, prior to identification of kinetically differentially expressed genes. The genome scale transcriptome data from wild type and a Δ*absA1* mutant of *Streptomyces coelicolor* were analyzed within this framework, and the regulatory elements affected by the gene knockout were identified. This methodology should find general applications in the analysis of other mutants in our repertoire and in other biological systems.

S. Mehra · W. Lian · K. Jayapal
S. Charaniya · W.-S. Hu (✉)
Department of Chemical Engineering and Materials Science,
University of Minnesota, 421 Washington Avenue SE,
Minneapolis, MN, 55455-0132 USA
E-mail: acre@cems.umn.edu
Tel.: +1-612-6267630
Fax: +1-612-6267246

D. H. Sherman
Life Sciences Institute, and Departments of Medicinal Chemistry,
Chemistry, Microbiology Immunology, University of Michigan,
210 Washtenaw Avenue, Ann Arbor, MI, 48109-2216 USA

Keywords Transcriptome analysis · DNA microarray · Streptomyces · Antibiotics · Time Series

Introduction

Differentiation of microorganisms gives rise to distinct morphologies under different growth environments and leads to the partitioning of structure and function in space and time. In higher organisms differentiation of different cell types results in the development of tissues, and eventually, in the formation of organs. The evolution of these time-dependent events is complex, entailing not only the alteration of cellular structure and morphology, but also the modulation or switching on or off of a large array of genes. In the past few years, DNA microarray analysis has become a powerful tool in examining the dynamics of these gene expression events. Of particular value is the study of gene expression profiles of mutants created by perturbing one or more genes that play a key role in controlling cellular differentiation. It is also interesting to study gene expression profiles of organisms growing under conditions that impose alterations in the differentiation process. Large-scale gene expression profiling of these mutants can potentially reveal genes of the same or related regulatory structure and further unveil the overall regulatory network. Since regulation is affected through diverse interactions of the genetic network, microarray-based transcription profiling conducted over a time span provides the information required to capture transcription dynamics.

Although DNA microarray analysis is becoming a common tool for gene expression exploration and the methodology for analyzing its data is largely developed, studies on time series of large-scale transcriptional profiles still face many challenges at both experimental and data analysis levels [3]. An accurate representation of the gene expression dynamics requires frequent sampling. Efficient experimental design and effective data treatment schemes are required to allow time series expression profiles across the different mutants to be compared. A major challenge in data analysis arises from the different rates at which growth and various developmental events unfold in different variants or strains of the same species or under different culture conditions. As a result, a direct comparison of transcriptomes between the same time points across different conditions might not be appropriate for the question being asked. In comparing different microbial strains, or mutants, one aims to compare the transcript levels in multiple dimensions, namely, transcript level of different genes at the same time point, different genes across different time points, and different gene sets across different time series data sets. In our study of the regulation of antibiotic synthesis, we faced the same multitude of issues and, thus, sought a solution to this problem.

Streptomyces species are soil-dwelling gram-positive bacteria that produce more than two-thirds of antibiotics in clinical use, as well as other natural products with diverse biological activities [2]. They differentiate into branched filamentous hyphae that give rise to aerial mycelia bearing long chains of reproductive spores. The formation of aerial mycelia and spores coincides with the onset of antibiotic production. In liquid culture, antibiotic biosynthesis typically follows the cessation of exponential growth. Antibiotic biosynthesis is considered secondary metabolism, because they are not essential for growth under laboratory conditions. It is believed that some secondary metabolites provide a competitive advantage in the microbe's natural habitat and their synthesis is among the most manifested example of events controlled by a biological clock that can be decoupled from growth.

Streptomyces coelicolor is the most studied species of the antibiotic producing *Streptomycetes*. Its 8.7 Mbp genome [4] encodes over 7,000 genes, greater than the number found in the lower eukaryote *Saccharomyces cerevisiae*. Its linear chromosome has more than 20 secondary metabolite gene clusters including those for the three antibiotic systems, actinorhodin, undecylprodigiosin and calcium dependent antibiotic [4, 12]. A significant fraction of *S. coelicolor* genes encode regulators or putative regulators, and many have been identified or implicated in the regulation of secondary metabolite biosynthesis [9]. Initial genome-wide gene expression profiling revealed remarkable dynamic behavior of a large number of genes, including those involved directly and indirectly in antibiotic biosynthesis and nutrient regulation [13]. Antibiotic biosynthesis is regulated at different levels [5]. Genes encoding enzymes for antibiotic biosynthesis are at the proximal level of the hierarchical structure and are under the control of pathway specific regulators [5]. Expression of the pathway-specific regulators is influenced by various local regulators that control the synthesis of specific antibiotics. On a more distal level [5], there is a layer of global regulators that pleiotropically control the production of more than one antibiotic.

An effective strategy to elucidate intricate regulatory networks is to genetically perturb key network components by constructing deletion mutants. Unlike genes involved in primary metabolism and other essential cellular functions, deletion of secondary metabolite biosynthesis and regulatory genes is non-lethal. Thus, mutants of nearly every regulatory gene involved in secondary metabolism can be identified and isolated, making it ideal for a systems biology approach to understand temporal and spatial genetic networks. With the availability of whole genome DNA microarrays it is appealing to examine the transcription profile of regulatory gene mutants, discern the impact on transcriptome dynamics and antibiotic production phenotype, and identify the genes affected by each mutation for reconstruction of the secondary metabolite regulatory network.

In this study, we focus on gene clusters of four secondary metabolite clusters: actinorhodin (ACT), undecylprodigiosin or Red (because of its red color), calcium-dependent antibiotic (CDA), and a putative type-I polyketide whose product has yet to be identified. We have constructed a DNA microarray to survey the transcription profile of various *S. coelicolor* regulatory gene disruption mutants. The mutants exhibit varying phenotypes including increased or decreased antibiotic production, and temporal shifts in production patterns amongst the different antibiotics. A first step to elucidate the regulatory structure of secondary metabolism is to identify genes that are differentially expressed in a mutant strain compared to the wild-type microorganisms. The dynamic behavior of gene expression demands analysis of the entire time series data set. We report here a systematic scheme to study those time series transcriptome profiles for identifying genes whose dynamics are altered in specific *S. coelicolor* regulatory mutant strains.

Materials and methods

Disruption mutant construction

The disruption mutants of *S. coelicolor* regulatory genes were created by replacing most of the coding sequence of the targeted gene with an apramycin resistance gene (*amr*). pDHS901, a derivative of the pGM160 *E. coli*/*Streptomyces* shuttle vector, was used to construct the disruption plasmid. Each of these plasmids contain a *tsr* gene, a 1 kbp fragment whose sequence is identical to the left-side flanking region of the target disruption area, the *amr* gene, and another 1 kbp fragment identical to the right side flanking region of the disruption area. All plasmid constructions were accomplished by four-way ligation, followed by transformation of *S. coelicolor* [15]. Using apramycin for selection allowed double crossover recombination between the plasmid and the *S. coelicolor* wild-type chromosome with concomitant loss of the plasmid. The putative disrupted clones were identified by the apramycin-resistant and thiostrepton-sensitive phenotype indicating gene disruption and plasmid loss, respectively. The clones were evaluated by PCR amplification of the target region of the *S. coelicolor* chromosome to confirm the identity of the desired mutation.

Culture Conditions

Batch cultures in liquid media were carried out as described previously [15], except as noted below. About 300 μl of a spore suspension (~10^{10} spores per ml) of *S. coelicolor* was inoculated into 40 ml of 2xYT in a 250 ml flask for pre-germination. The flasks were incubated at 30°C on a shaker at 300 rpm for 6–8 h. Samples were taken periodically to check for presence of emerging cells. Once visible, the cell suspension was spun down at > 1500 × *g* in a 50-ml centrifuge tube, washed with modified R5 medium [13] and resuspended in 5–10 ml of modified R5 medium. The pre-germinated spores were inoculated into 2-l flasks containing 300 ml of the main culture medium with an inoculum concentration ~10^7 spores per ml. The flasks were incubated at 30°C on a shaker at 300 rpm as before. Samples were taken periodically to measure growth and antibiotic titers and for RNA extraction. For measuring growth, 0.5 ml of sample was diluted with 400 μl water and 100 μl 2.5N HCl [7] and was dispersed by sonication for 30–60 s. The dispersed suspension was then diluted to a desired level and absorbance at 450 nm was observed. The antibiotic titers were also measured spectrophotometrically as described [8, 15]

cDNA Microarray Construction

A whole genome cDNA microarray was constructed for *S. coelicolor*. Gene-specific probes with an average length of 500 bp were designed based on the genomic sequence database at Sanger Center to represent the 7,825 genes [4]. The primers to amplify these probes were designed to be 20 bp in length with a melting temperature around 62°C. An iterative primer/probe design process was adopted. The primers were first Blasted against the entire genome to check for the specificity to the corresponding genes. The probe sequences resulting from these primers were also blasted against the entire genome to check for possible cross-hybridization with sequences other than the corresponding gene. Any primer pair that may give rise to a potentially croshybridizing probe sequence was redesigned with alternate coordinates. This probe checking-redesign process minimized the possibility of cross-hybridization between a probe and other cDNA species. For some highly homologous genes, a unique probe was not attainable and a multigene probe was used. A multigene probe was designed in such cases. 7,581 genes were successfully amplified and spotted on the microarray. The microarray was fabricated on poly-L-lysine slides with duplicate spots for each probe.

Genomic DNA Reference preparation

Genomic DNA was isolated from mycelia at the stationary growth stage in YEME liquid culture following the Kirby mixture method [15]. Genomic reference DNA for the microarray analysis was prepared by fragmenting the isolated gDNA to 500 to 2 kbp average range. A nebulizer with 1 mg of gDNA in 2 ml of a DNA buffer containing 40% glycerol was placed in an ice-bath and was subjected to compressed nitrogen gas at a pressure of 25 psi for 3 min. The resulting DNA fragments were purified by ethanol precipitation and resuspended to a concentration of about 1 μg/ μl. The fragmented genomic DNA was then labeled with Cy3 dye using *Label* IT® Cy™3 Labeling Kit (Mirus, Madison, WI, USA). The labeling reaction consisted of 20% Label IT® Reagent and 1 μg of genomic DNA in a 7 μl reaction volume. The reaction was incubated at 37°C for 3 h and the labeled genomic DNA was purified with a MinElute PCR purification kit (Qiagen Inc., Valencia, CA, USA) as per the manufacturer's instructions.

RNA probe preparation

Total RNA was isolated by fragmenting mycelia with mortar and pestle in the presence of liquid nitrogen and subsequent purification using RNeasy Mini Kit (Qiagen Inc.). Total RNA was reverse transcribed to cDNA with concomitant incorporation of aa-dUTP (Ambion, Austin, TX, USA) using random hexamer primers and Superscript™ II reverse transcriptase (Invitrogen, Carlsbad, CA) at 42°C for 2 h. The cDNA was then labeled with Alexa 647 (Invitrogen, Carlsbad, CA, USA).

Microarray hybridization

For each microarray hybridization, 200 ng of genomic DNA (gDNA) reference and 10 μg of total sample RNA were used. All hybridizations were carried out in a buffer containing 50% formamide at 50°C for 16 h. Details of all protocols are described at https://hugroup.cems.umn.edu/Protocols/protocol.htm.

Image quantification and data analysis

The hybridized slides were scanned using ScanArray 5000 scanner (Perkin Elmers, Boston, MA, USA) and the images were analyzed using Genepix (Axon Instrument, Union City, CA, USA). The Genepix data (.gpr) files were exported and further analyzed using an inhouse written code in the Matlab computing environment. Spots with a small diameter, low intensity in the genomic DNA channel, or those that were flagged during image analysis were filtered out before further analysis. For each spot, a log ratio to the base 2 was calculated between the intensity of cDNA to gDNA reference. This ratio is henceforth referred to as the log2 expression value. The spots were normalized using either linear or quantile normalization. For quantile normalization, the log ratios from the array to be normalized and the reference are sorted and every percentile value from one array is compared against the other. To normalize the data, any percentile value from the given array is replaced by its corresponding value from the reference array. For log2 ratios that fall between any two percentile values, a cubic spline fit interpolation of the data is used. The normalized data is then unsorted back to the original order to complete the process. The log2 expression values of the replicate spots corresponding to the same gene were averaged, and a standard deviation calculated. Any spot that has a log2 ratio out of the bound of 1.2 times the standard deviation was removed as an outlier. For the rest of the spots, an average and standard deviation was recalculated. The complete normalization process has been automated in the Matlab computing environment. The program takes Genepix or Imagene raw data files as input and gives the normalized averaged data as output.

Real-time PCR Analysis

Reverse transcription (RT) of total RNA followed the same procedure as for microarray probe preparation. Forward and reverse primers were designed by Primer3 (http://frodo.wi.mit.edu/) and agarose gel electrophoresis was used to confirm the amplicon size. QuantiTect SYBR® Green PCR kit (Qiagen) was used to perform real-time PCR on ABI7700 (Applied Biosystems, Foster City, CA, USA). 12 µl of the reaction mixture contained 6 µl SYBR® Green PCR Master Mix (Applied Biosystems), cDNA template from 10 ng RNA, 250 nM each of forward and reverse primers and MilliQ water. The PCR reaction was initiated by incubating the sample at 95°C for 15 min followed by 40 cycles at 94°C for 20 s, 60°C for 30 s, and 72°C for 15 s. To check the specificity of the real-time PCR reaction, a DNA melting curve analysis was performed by holding the sample at 60°C for 60 s followed by slow ramping of the temperature to 95°C. A threshold of ten times the average standard was used as a cut-off for C_t (threshold cycle number). The specificity of the PCR product was confirmed from a single peak in the dissociation curve. SCO4548 gene was used to normalize across RNA samples from different time points. For each gene, the expression at time point t was calculated as a log2 ratio with respect to the 18 h time point as

$$\frac{2^{C_{\mathrm{T}}(\mathrm{gene},t)-C_{\mathrm{T}}(SCO4568,t)}}{2^{C_{\mathrm{T}}(gene,18\,h)-C_{\mathrm{T}}(SCO4568,18\,h)}}.$$

Dynamic time warping for time series alignment

The time warping algorithm uses dynamic programming to minimize the distance between two G-dimensional time series $X(u)$ (reference) and $Y(v)$ where $u = u_1, ..., \mathrm{u}_{\mathrm{N}}$ and $v = v_1, ..., v_M$ are N and M time points, respectively, where the series are sampled. The two series are represented on a $N \times M$ grid, where the vertical and horizontal axes of the grid (index i and j) correspond to the time points of the X and Y series, respectively. The distance between the ith time point on X and jth timepoint on Y, $D(i,j)$ is found recursively as

$$D(i,j) = \min \begin{cases} D(i-1,j) + \frac{(u_i - u_{i-1})}{2}\frac{d[i-1,j]+d[i,j]}{2} \\ D(i-1,j-1) + \frac{(u_i - u_{i-1} + v_j - v_{j-1})}{2}\frac{d[i-1,j-1]+d[i,j]}{2} \\ D(i,j-1) + \frac{(v_j - v_{j-1})}{2}\frac{d[i,j-1]+d[i,j]}{2} \end{cases}$$

where $D(1,1)=0$, $d[i,j] = \sqrt{\sum_{l=1}^{G} f_l(X_l(u_i) - Y_l(v_j))}$ where f_l is the weight assigned to each dimension and $D(i,j)=\infty$ if $|i-j| < T_{\max}$. where $T_{\max}$ is the maximum shift allowed. The grid point that leads to the minimum $D(i,j)$ is stored. The optimal path is constructed by tracing the path from $D(N,M)$ to reach $D(1,1)$. Once the optimal path is determined, profile Y is mapped to the time scale of X. If multiple time points of Y correspond to a single time point of X, the expression value in the new Y series, is an average over the multiple time points.

Results and discussion

Genomic DNA as a common reference for hybridization

The growth and antibiotic production kinetics of *S. coelicolor* wild type and a mutant strain, Δ*absA1*, where the sensor kinase of the corresponding two-component system has been disrupted, are shown in Fig. 1. The growth curves for mutant and wild type exhibit the characteristic exponential and stationary growth phase. In wild type the appearance of antibiotic *Red* coincided with the transition from rapid exponential growth phase to stationary phase. The production of *Red* in Δ*absA1* was delayed compared to the wild type and accumulated to a lower level. The antibiotic *Act*, produced during late stationary phase, also appeared earlier and accumulated to a higher level in the wild type as compared to the mutant culture. Cell samples from different time points during the culture were taken for transcript profiling using the whole genome DNA microarrays.

Gene expression at the transcript level exhibits very fast dynamics in *S. coelicolor*. The transcript level of some genes reaches its maximum value and subsides within a few hours. To capture the dynamics of the expression of these genes, frequent sampling during the critical period of cultivation is required. The number of samples is further expanded by the large number of potential knockout mutants of known and putative regulatory genes already or yet to be created. The total number of microarray hybridizations that need to be performed can become prohibitively large.

In a traditional two-dye microarray, transcript analysis is performed by pairing the samples. The transcript levels in two samples are compared to generate a ratio of one sample versus the other. Ideally, the experimental

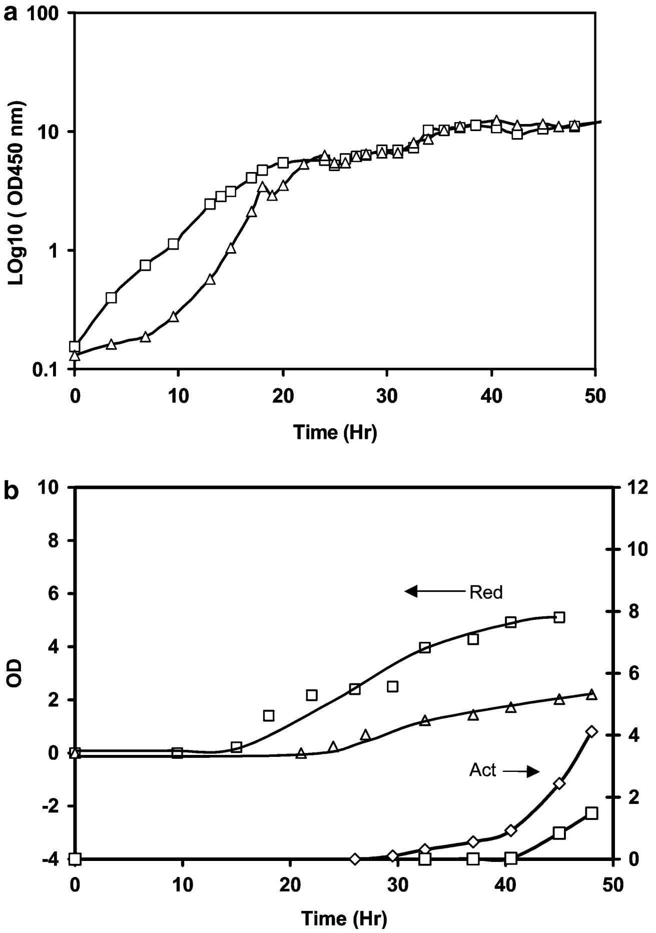

Fig. 1 Growth and antibiotic profiles for wild type *S. coelicolor* and Δ*absA1* strain. **a** Growth curve **b** the Red and Act antibiotic profiles. *Open squares* correspond to wild-type *Streptomyces coelicolor* and *triangles* are for the Δ*absA1* strain. The sampling points for optical density (O.D.) and antibiotic measurements are shown. Note the increase in sampling frequency around the Red onset times

design of pairing of samples should allow for cross-culture and cross-mutant comparison. A sample at a fixed time point, typically the first sample, is frequently used as a reference, allowing all other samples in the same culture to have a common reference. However, many genes like those involved in secondary metabolism and differentiation are not expressed at the reference time point and therefore are not represented in the reference sample. The resulting low fluorescence intensities from these genes introduce inaccuracies in the corresponding ratios. When multiple cultures are to be compared, as in the case of comparative studies of different knockout mutants, the need for a common reference is not only for within a culture, but also across different cultures or mutants. One approach is to pool "representative" RNA from a large number of samples as a reference. To maintain the consistency of the reference sample, a large amount of RNA from many samples needs to be isolated. With the large number of potential mutants to be generated over time, this approach becomes cumbersome and impractical.

We have taken the approach of using genomic DNA (gDNA) as a common reference for all hybridizations. Successful use of gDNA standards in microarray hybridizations has been demonstrated earlier for *Mycobacterium tuberculosis* and mammalian cells [20, 21]. The ease of extraction, reproducibility and relative invariance of gDNA make it an ideal candidate for use as a reference standard. However, one potential concern while using gDNA reference is that, unlike single strand cDNA synthesized by reverse transcription from RNA, gDNA is double stranded. Hybridizations involving gDNA are thus different from conventional cDNA:cDNA hybridizations. The immobilized probe on the microarray surface competes for available cDNA with the complementary strand of gDNA in solution. To alleviate this concern, we have developed a mathematical model and assessed the degree of error in theoretical expression ratios introduced due to solution phase hybridization under different operating conditions [11]. The analysis using the model shows that such interference is relatively small under most operating conditions.

An approach to overcome the requirement of a common reference is to use a loop design in a two-dye microarray hybridization [14]. It is a combinatorial approach encompassing all possible or a minimum set of pairs of samples. The minimum set is a closed loop pairing every two consecutive samples in sequence. The denominator sample in a pair becomes the numerator sample in the next pair. The combinatorial experimental design requires a large amount of RNA for each sample; while the ratios calculated using a loop design have variable levels of precision. This is because the ratios for some samples are directly obtained from a single hybridization, while others are indirectly computed from multiple hybridizations. Other disadvantages of the loop design include the inconvenience of inserting a new sample into the earlier 'loop' and the peril of breaking up the loop if one or more paired hybridizations fail.

Normalization

Employing genomic DNA as a standard reference for two-dye microarray hybridizations necessitates a fresh look at data normalization. With the exception of a few, all genes are equally represented in genomic DNA, while the levels of transcript of individual genes vary over a few orders of magnitude. In microarray hybridizations, this gives rise to a relatively narrow range of reference (gDNA) channel intensities (G) and a wide distribution of sample (cDNA) channel intensities (R). Nonlinear normalization techniques like Lowess [22] are based on the premise that changes are approximately symmetric at all intensities and therefore a $\log(R/G)$ vs. $\log(R{*}G)^{1/2}$ plot should be centered around a log2 ratio of 0. These are therefore inadequate for processing intensity data obtained from hybridizations of genomic DNA versus cDNA comparisons.

A simple normalization method that can be applied for such circumstances is a linear, intensity based normalization. The same concentration of genomic DNA and cDNA is used in all experiments. Therefore, the total intensities of reference and sample channels should be constant across different hybridizations. Any deviation of the total intensities from the constant value is attributed to experimental perturbation and can be corrected by proportional scaling of all spot intensities for each channel. This normalization method is reasonable when the ranges of observed gene expression values are approximately the same in all experiments. However, nonlinear variations in intensities may arise from a number of factors, including microarray preparation, RNA quality, reverse transcription efficiency, and laser energy level in microarray scanning. The experimentally observed relative intensity between the two channels, even after linear normalization, at times exhibit different distribution. Figure 2a shows one example with samples taken from six different time points of a culture where such variations appear particularly prominent. It can be seen that the range of relative intensity between two channels is significantly smaller at 24 and 42 h compared with the rest of the samples.

To negate these presumptive irregularities, we employed a *quantile normalization* method [6]. The premise is that the overall distribution of total mRNA remains relatively constant across different samples. In this method, the log expression values from all hybridizations are normalized to match a single reference distribution function. We employed a reference distribution that is the average of all historical data. The algorithm for quantile normalization is described in detail in Materials and methods. Data obtained after quantile normalization is shown in Fig. 2b. The narrower range of intensity distribution at 24 and 42 h seen after linear normalization is compensated after quantile normalization. For all samples, the bounds of gene expression values are approximately the same after this normalization.

The transcript levels for different genes obtained from two-dye microarray hybridization are relative values, normally expressed as a logarithm to the base 2 of sample relative to the reference. When gDNA is used as a common reference, the relative expression of a particular gene between two samples is obtained by taking the ratio of their log2 expression value (both relative to gDNA). This preserves uniform accuracy across all ratios. When gDNA is used as a common reference, this ratio is an arbitrary number indicating the abundance of transcripts for a particular gene with respect to gDNA. The value is dependent on the amount of gDNA used in the assay. We therefore refer to the transcript level relative to the gDNA as the log2 expression value.

In Fig. 2c, d linear and quantile normalizations are compared for two replicate hybridizations from the 24-h sample. A plot of log2 expression values from the two hybridizations should collapse to a 45° line. In the case of linear normalization, a skewed distribution of the data points is apparent. After quantile normalization, the data indeed follow a 45° line (Fig. 2d).

Representative genes with their intensity values spanning a wide range were selected for quantitative Real-Time PCR analysis for comparison with microarray data. QRT-PCR was performed for RNA samples from 18, 28, and 42 h for Δ*absA1* strain. Relative expression levels of the genes were calculated at 28 and 42 h, with respect to the 18-h sample. The ratios were compared to those obtained from microarray after linear and quantile normalization, as shown in Fig. 3. The log2 ratios obtained after quantile normalization are all close to the real-time PCR values.

A number of causes can contribute to the apparent “compression” of gene expression. Furthermore, the effect on hybridization is likely to be nonlinear. Quantile normalization may not correctly “restore” the transcript level of all genes to their true value. Nevertheless, the results are satisfactory, and the method allowed most data to be analyzed and genes of potential interest to be identified for further investigation. Caution still needs to be taken in drawing firm conclusions.

Visualization of microarray data

Since a common genomic DNA reference was used in all microarray hybridizations, the expression level of all genes in a given sample, as well as across different samples from various time points of cultures with different mutants can be directly compared. It enables direct comparison of the dynamics of gene expression time profiles across different mutants. As each regulatory gene affects the expression of a multitude of genes under its direct or indirect influence, the ability to compare the effect of a specific mutation on a large number of genes is important. Figure 4 shows the relative expression value of transcripts for *S. coelicolor* wild type and Δ*absA1* across the entire genome at two time points (10 and 37 h). The first time point is during the exponential growth phase, while the second corresponds to the late stationary phase. Each data point on the chart represents the moving median transcript level of five neighboring genes at the corresponding position on the linear chromosome. The log2 expression values are shown. Genes with intensity below the reference gDNA channel have a negative log2 expression value.

As can be seen in Fig. 4, genes expressed at high levels are located closer to the middle region rather than at the two ends of the linear chromosome. A region rich in highly expressed genes lies between *SCO4700* and *SCO5700*. Ribosomal genes, clustering near *SCO4700*, and ATP synthase genes, clustering around *SCO5370*, are among those with highest expression levels in this region. The expression levels of these genes subsided somewhat at 37 h as compared to 10 h but remain relatively high in both strains. It is notable that the antibiotic cluster *Act* (around SCO5080) is expressed at 37 h and not at 10 h in the wild type.

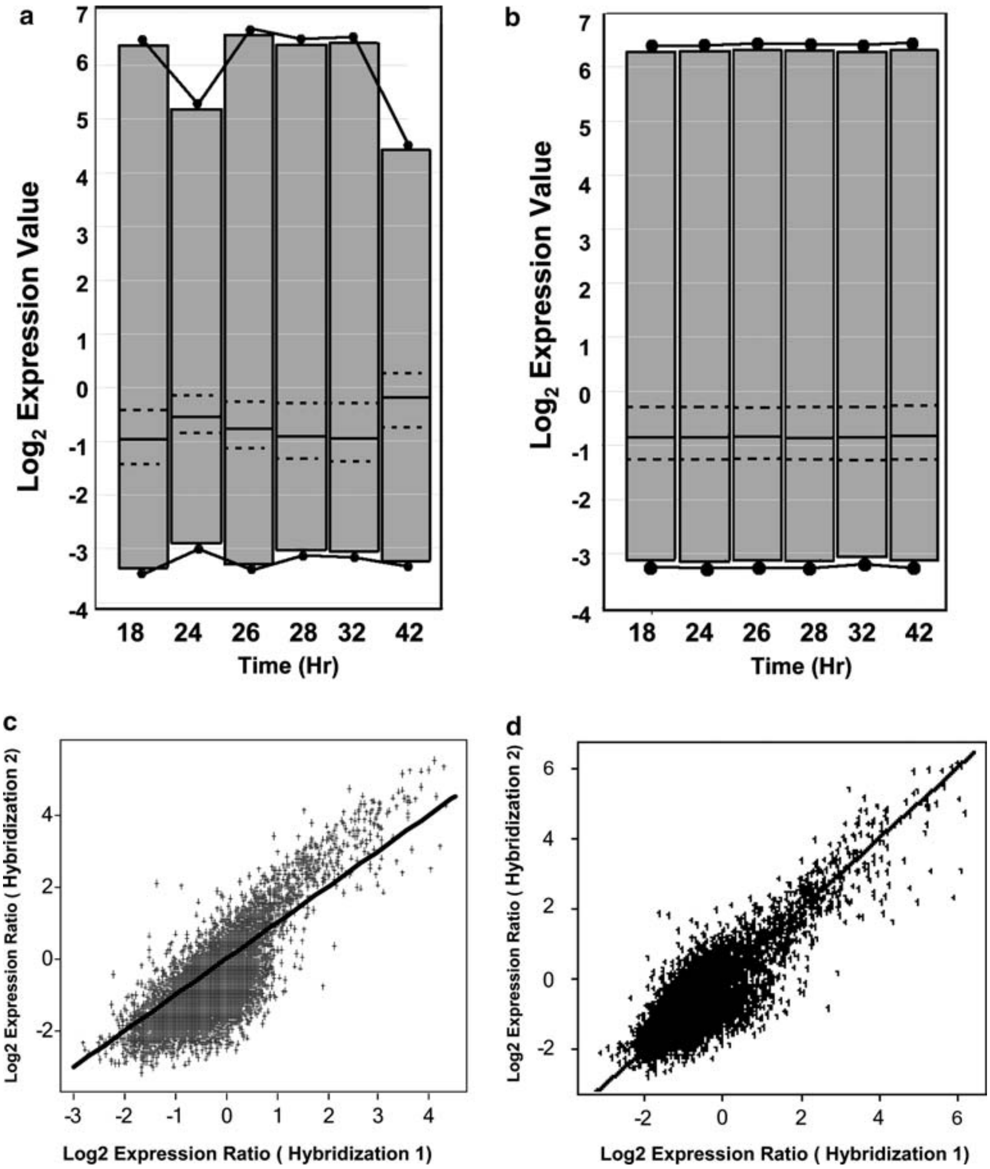

Fig. 2 Comparison of linear and quantile normalized time-profile data obtained from microarray analysis. **a** and **b** Box plots showing the bounds of log expression ratios for samples taken at different growth stages in a batch culture of Δ*absA1* mutant strain. The centerline within each box corresponds to the median value, while the two dotted lines above and below are the 75th and 25th percentile respectively. The upper and lower bounds of the box are the maximum and minimum log ratios for each sample. **a** Data obtained after intensity-based linear normalization, **b** the quantile normalized form of the same data. Log2 expression ratios of all genes from 24-h sample are plotted against a replicate hybridization after **c** linear and **d** quantile normalization

A comparison of the transcript profiles of the two strains at 37 h shows that their differences are distributed along the entire chromosome. The most prominent divergence includes the antibiotic cluster of *Act*, which is expressed at a conspicuously high level in wild type, but is almost entirely suppressed in Δ*absA1* as a result of the delayed onset of Act biosynthesis. A similar delay in onset times of the other three clusters (*Red, CDA*, type I polyketide) is also observed. Such alteration in gene expression profiles may be a genuine effect of the genetic mutation. Conversely, these temporal shifts may be the results of culture conditions or even stochastic in nature, as occasionally seen even in biological replicates.

Time shift of gene expression profile, thus, poses a challenge for identifying those genes truly altered by the mutation. If a direct comparison of expression profiles for mutant and the wild type is performed at the same time point, genes whose expression have shifted in time would be identified as having an altered kinetic behavior. However, the observed change in gene expression may be attributed to a longer lag phase or a different growth rate of the mutant. The gene expression profiles of different cultures or different mutants therefore need to be aligned to those of the wild type at similar stages of growth or development before comparing the data to identify differentially expressed genes.

Direct genome-wide visualization of transcript levels highlights the global positional effects of the mutation. However, such a comparison does not identify individual genes whose transcriptional kinetics have changed. Analyzing gene expression at the functional class and individual gene level can reveal the complete effect of the knockout mutation. It allows us to identify genes whose expression level or dynamics has been altered between a

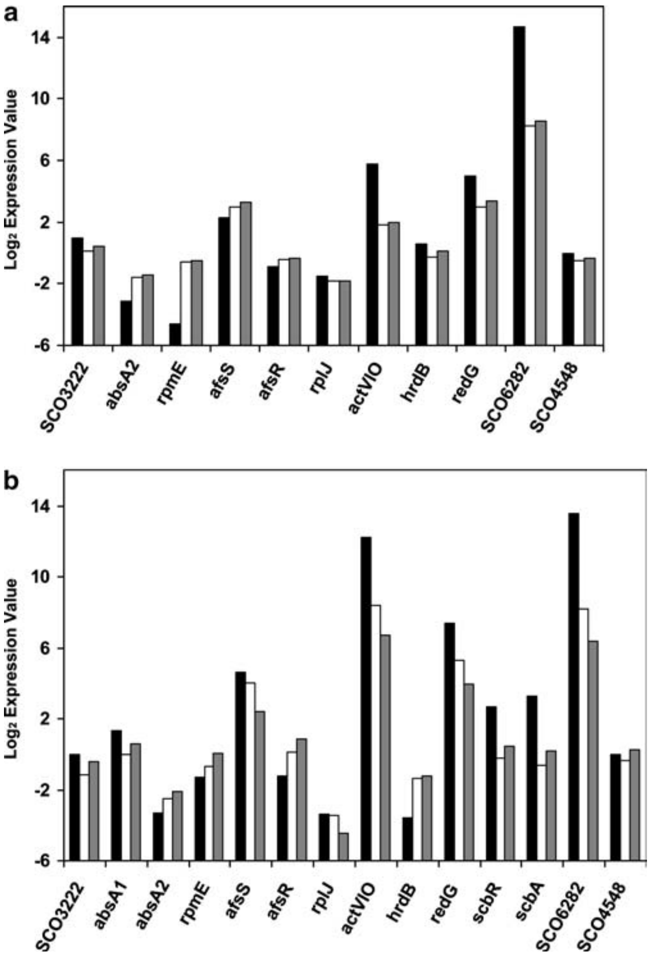

Fig. 3 Comparison of expression ratios from microarray analysis with quantitative real-time PCR. Log2 ratio of genes from samples at **a** 28 h and **b** 42 h were obtained with respect to 18 h. The three bars correspond to real-time PCR data (*black*), linear normalized microarray data (*gray*) and quantile normalized microarray data (*white*)

mutant and the wild type at any time point. These genes are likely to be under the direct or indirect influence of the disrupted gene.

Time alignment

Growth and differentiation events in developing organisms are under the control of multiple biological clocks. These different clocks and the corresponding set of genes may respond differently to various environmental or genetic perturbations. Figure 5 illustrates several scenarios where the time profiles of growth and gene expression may respond to a genetic or environmental perturbation, causing them to be asynchronized with the reference culture. Globally, when a lag phase arises, the expression profiles of most genes are linearly shifted with respect to the reference profiles, herein referred as frame shift. Conversely, a higher or lower growth rate with respect to the reference will be manifested as an elastically compressed or stretched profile. A situation may arise where only one of the local clocks is affected causing a possible time shift in expression profile of a set of genes. This may result in a change in time order of gene expression with respect to another class of genes that follows a different clock. Before a comparison between two strains can be made, the gene expression profiles should be aligned with the reference culture such that similar growth stages correspond.

The alignment of the growth and gene expression profiles of two cultures can possibly be undertaken at two levels—global and local. Alignment at a global level takes the time profiles of all the transcripts into consideration to find an ‘‘optimal’’ relative time scale for the target culture with respect to the reference culture; whereas at a local level one may consider only a subset of genes and align only those genes from the two cultures. An example for the latter case is the alignment of *Red* or *Act* genes from mutant and wild type. For mutants with genetic alteration(s) in secondary metabolism the kinetics of growth and nutrient consumption and the progression of changes in environmental factors are likely to be relatively invariant. A mutant may have a different lag phase, or different growth rate, but on the global level the sequence of events in both does not reverse. We can thus assume that the vast majority of genes preserve a similar expression profile as in the wild type. However, at the local level a subset of genes would have an altered expression profile as a direct or indirect result of the gene mutation. This is supported by the results shown in Figure 4; a relatively small proportion of transcripts change their expression profiles as a result of the knockout.

We address the issue of alignment using dynamic time warping (DTW), a technique originally developed for speech recognition [17], to recognize words spoken by different speakers with varying duration and intensities. The objective of the method is to match two patterns, sequences or trajectories by locally translating, expanding, or compressing patterns such that similar characteristics within the patterns are aligned. The method has been widely used in a variety of fields, including synchronization of batch cultures for fault detection etc. Recently, based on this algorithm Church et al. [1] have demonstrated its use to align cell-cycle data.

The implementation of this algorithm is described in Materials and methods. The gene expression profiles of two strains can be considered as two *G*-dimensional trajectories, $X(u)$ and $Y(v)$, where *G* is the number of genes, and, $u = u_1, ..., u_N$ and $v = v_1, ..., v_M$ refer to the *N* and *M* sample time points respectively. Thus, $X(u_1)$ refers to a *G*-dimensional vector representing the expression value of all genes at time point u_1. The time warping algorithm maps the time indices of the two series to a common index, such that the distance between the two series is minimized. The new time index consists of *K* points where each point is an ordered pair $(i(k), j(k))$, such that $u_{i(k)}$ corresponds to $v_{j(k)}$. Multiple time points of one series are allowed to correspond to a single time point of the other series to allow for compression and expansions. K is bound by

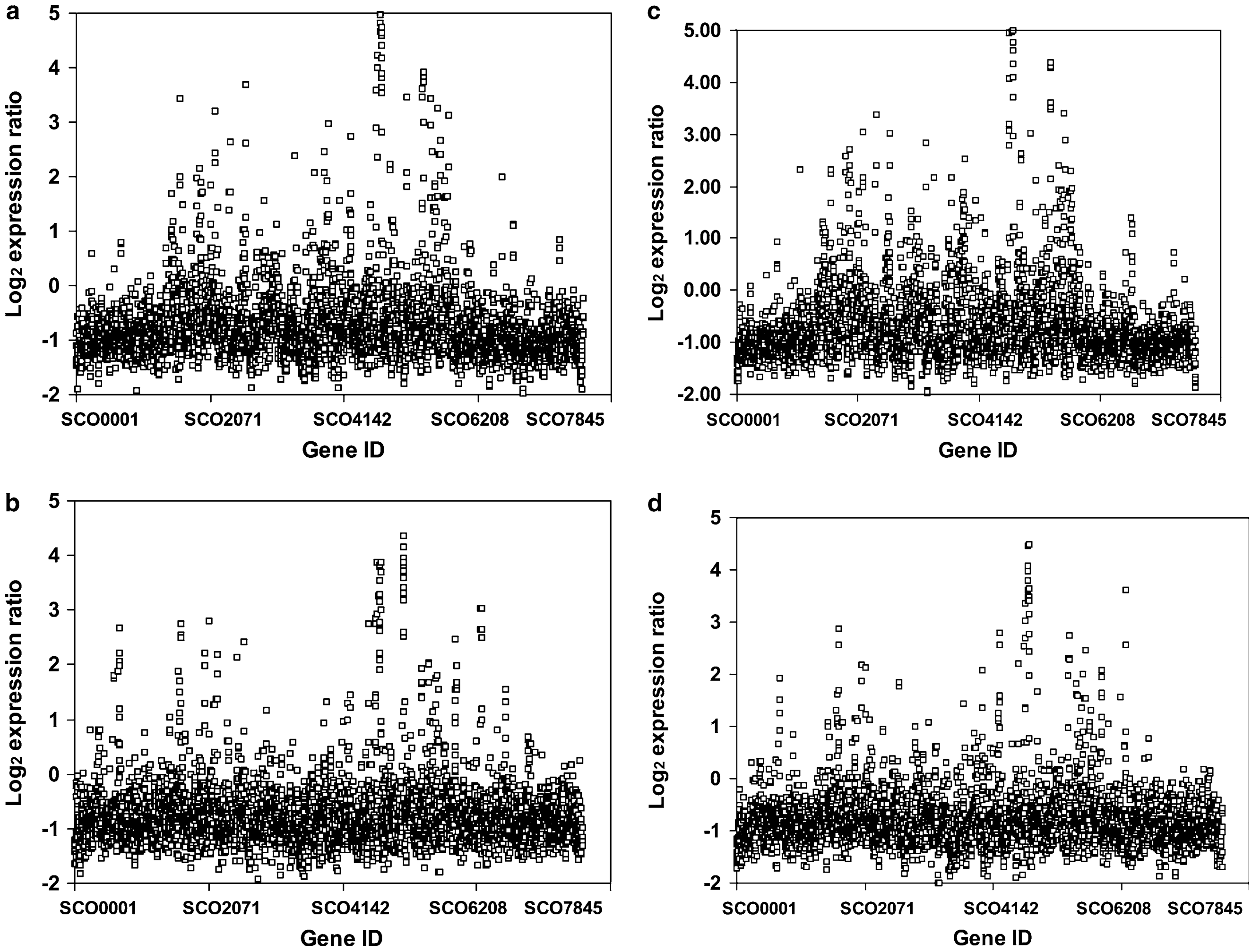

Fig. 4 Relative expression level of cell transcripts across the chromosome for wild type and Δ*absA1* mutant at two different times along the growth curve. **a**, **b** 10 h and 37 h for the wild type, **c**, **d** transcript levels from 10 h and 37 h of Δ*absA1* mutant. Each point is the moving median of five adjacent genes on the chromosome

maximum($n + m$) ≤ $K \leq n + m$. The distance between time series X and Y is defined as follows:

$$D(X,Y) = \frac{\sum_{k=1}^{K} d[u_{i(k)}, v_{j(k)}].w(k)}{N(w)},$$

where $d[u_{i(k)}, v_{j(k)}]$ is the distance between $X(u_i)$ and $Y(v_j)$.

$$w(k) = \frac{(u_{i(k)} - u_{i(k-1)} + v_{j(k)} - v_{j(k-1)})}{2}$$

$$\text{and } N(w) = \sum_{k=1}^{K} w(k) = \frac{(u_n - u_1 + v_m - v_1)}{2}.$$

$N(w)$ is a normalization factor such that the distance is independent of the length of the number of path points K and the length of the two trajectories. The distance measure $d[u_{i(k)}, v_{j(k)}]$ can be the Euclidean or Pearson correlation, depending on the purpose. Additionally, each gene maybe weighted differently to compute the distance measure. To ensure order and continuity, minimization of this distance is subject to the following constraints.

1. Boundary conditions

(a) $i(0) = u_1$ and $j(0) = v_1$.
(b) $i(K) = u_n$ and $j(K) = v_m$.

2. Continuity.

(a) $i(k) - i(k\text{-}1) \leq 1$ and $j(k) - j(k\text{-}1) \leq 1$. This forces the warping path to move to adjacent or diagonal cells.

3. Monotonicity

(a) $i(k) - i(k\text{-}1) \geq 0$ and $j(k) - j(k\text{-}1) \geq 0$. This preserves the order of events.

4. Global Bound

(a) $i(k) - j(k) \leq T_{\max}$, where $T_{\max}$ is the maximum allowed shift of one series with respect to another.

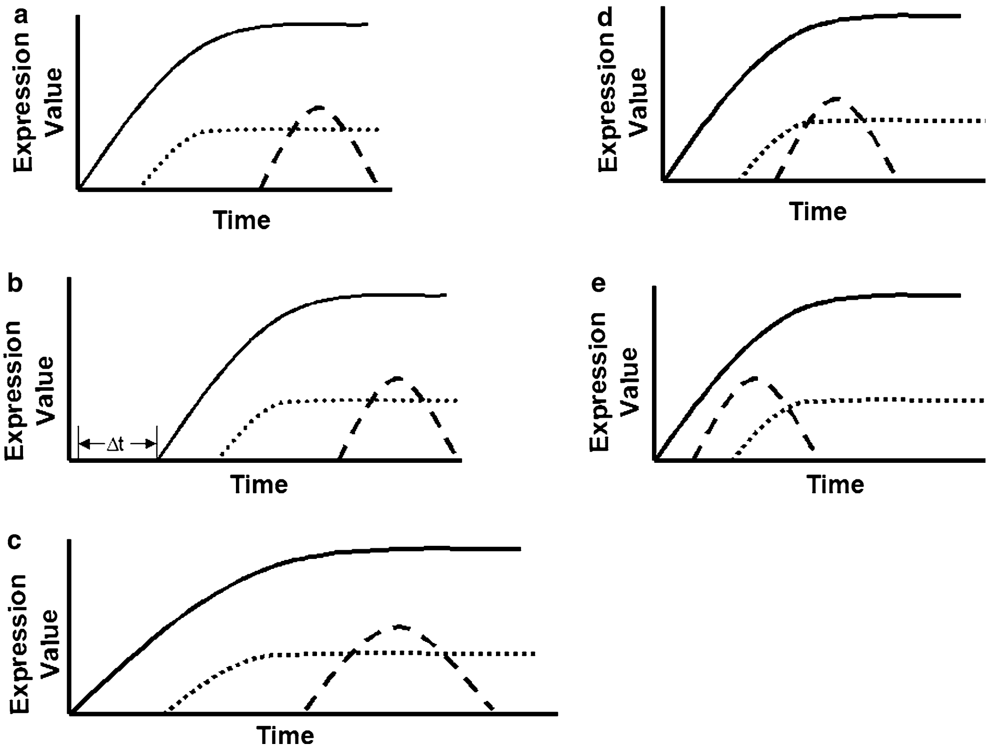

Fig. 5 Possible variations in genotypic and phenotypic dynamics between different cultures. reference culture (**a**), frame shift (**b**), elastic expansion (**c**), elastic compression (**d**) and time flip (**e**). The three curves correspond to growth (*solid line*), gene 1 (*dotted line*) and gene 2 (*dashed line*)

Employing this algorithm, the transcriptome time profiles from *S. coelicolor* wild type and mutant strains were aligned. The Euclidean distance measure was used to compute an optimal alignment between the two sets of time series data. The time-alignment algorithm needs a basal set of genes exhibiting profound dynamics. To improve alignment stability, genes that show a relatively static profile (where the range of expression value is less than twofold) were removed. The remaining gene set was used to compute a global optimal alignment path.

Figure 6 shows the global time alignment path between the time profiles of genes from the wild type and Δ*absA1* mutant. The absolute time points for the two strains are on the *x*- axis and *y*-axis, respectively. The 45° line represents a one-to-one correspondence and a perfect match between the time points. The alignment between wild type and Δ*absA1* mutant lies above the 45° line, demonstrating that the Δ*absA1* culture lags behind the wild type. This is reflected in the frameshift alignment during the early growth stage (until 23 h on the *y*-axis), where a later time point from Δ*absA1* corresponds to an earlier time point from wild type. After 23 h, a transition in the alignment path is followed by another frameshift in the Δ*absA1* mutant of a lesser magnitude.

Based on the alignment path, the expression profiles from the mutant are projected on to the time scale of the wild type, as described in Materials and Methods. Figure 7 shows two antibiotic clusters before and after alignment of the Δ*absA1* mutant and wild type. The onset of *Red* and *Act* clusters were 16 and 27 h and were 24 and 37 h in wild type and Δ*absA1*, respectively. Without alignment, the *Red* and *Act* antibiotic cluster genes in the two strains show major differences in their expression profiles. After alignment no such difference was apparent in their kinetics. In contrast, a set of genes involved in sulfur metabolism was delayed in their peak expression in the wild type as compared to the mutant strain. Note that the relative order of this cluster of genes in the two strains was reversed after alignment.

The above implementation of alignment of time-series profiles of the wild type and *absA1* mutant assumes that the sampling rate does not introduce any bias, and time points from the wild type can be paired with a data point from the mutant profiles. If the sampling time-points from the two cultures are not equivalent, an interpolative time-alignment algorithm can be implemented. In such a case, the time points from any one

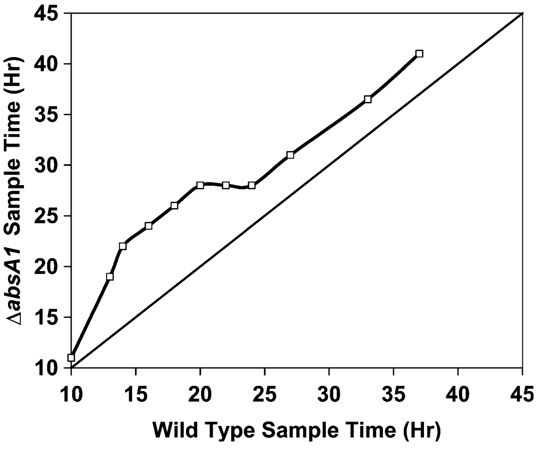

Fig. 6 Global alignment path between wild type and Δ*absA1*. The abscissa and ordinate correspond to time points from the wild type and the ΔabsA1, respectively

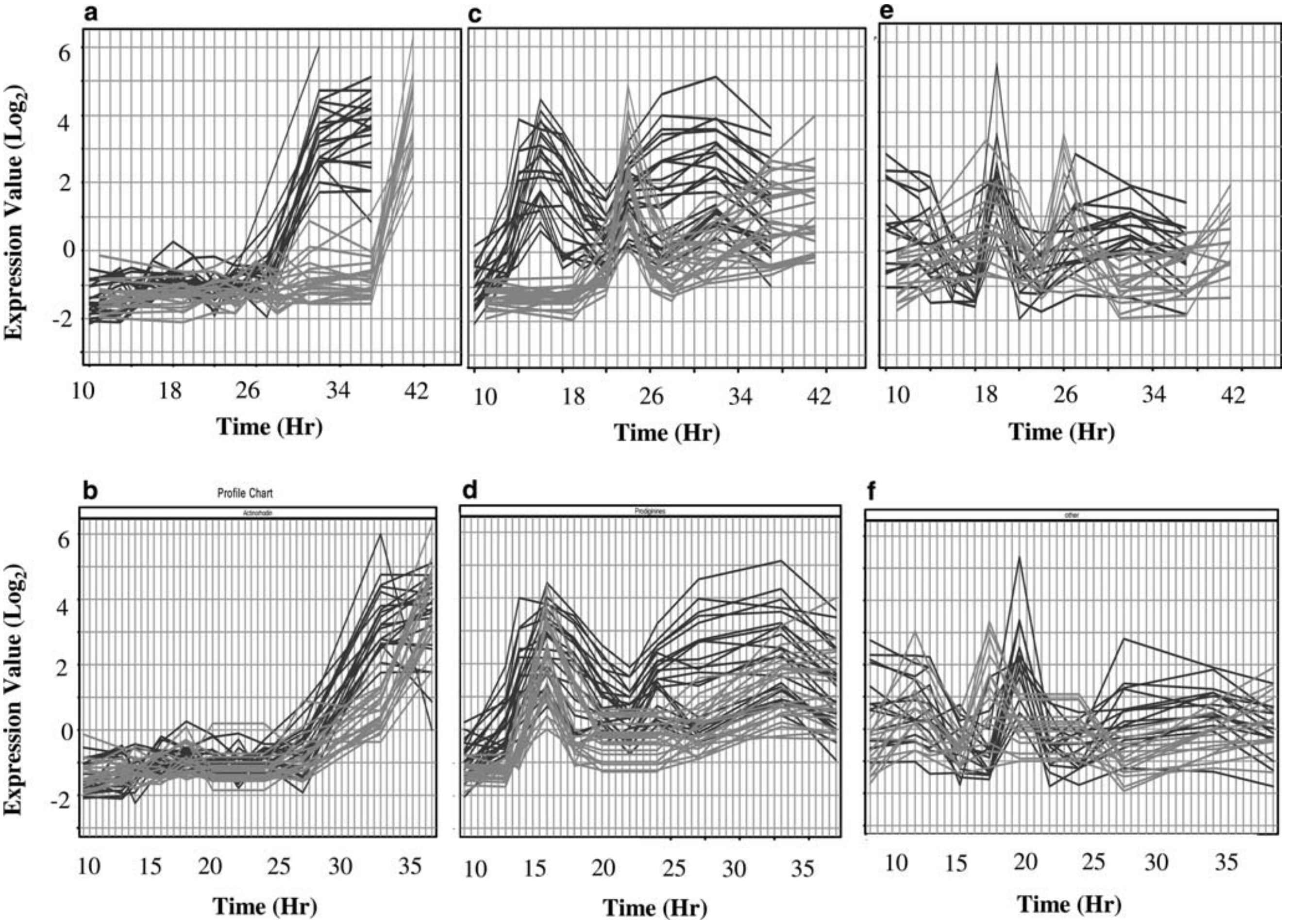

Fig. 7 Effect of time alignment on the expression profiles of three gene clusters in the wild-type strain (red) and Δ*absA1* (green). The top panels correspond to the data before alignment and the bottom panels illustrate the profiles after alignment. **a**, **b** the Act antibiotic cluster, **c**, **d** the Red antibiotic cluster, and **e**, **f** the profiles for a group of genes involved in Sulfur metabolism

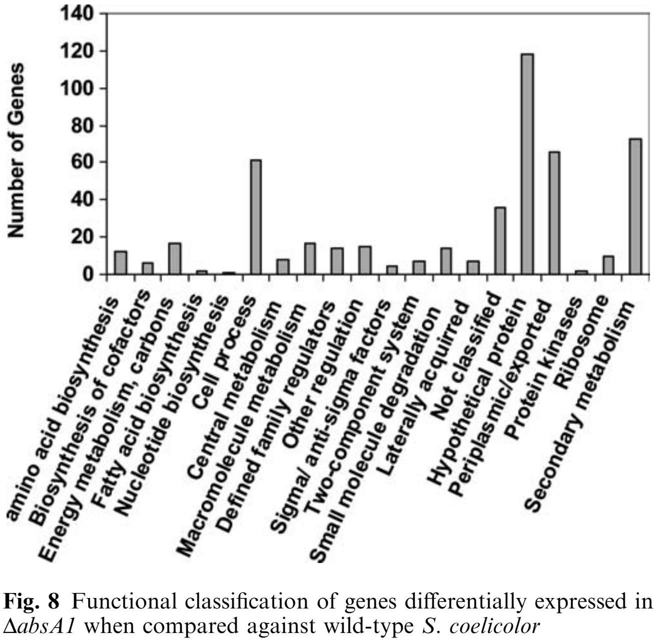

Fig. 8 Functional classification of genes differentially expressed in Δ*absA1* when compared against wild-type *S. coelicolor*

culture are allowed to be mapped to an interpolated time point of the other culture and vice versa. For the time profiles of the wild type and Δ*absA1*, the interpolative implementation resulted in a similar alignment path as the discrete algorithm present in this work (data not shown).

Identification of kinetically differentially expressed genes

After aligning the time series expression profiles, a number of statistical measures are used to identify differentially expressed genes. The Euclidean distance between the log2 expression profile of each gene in the wild type and mutant was used as a criterion of differential expression. However, genes with statistically insignificant differential expression at multiple time points may also give rise to a high Euclidean distance. Furthermore, Euclidean distance is insensitive to random fluctuations of one series against the other. Therefore, genes with less than 1.4-fold ratio of one strain as compared to the other at all time points were eliminated. In addition,

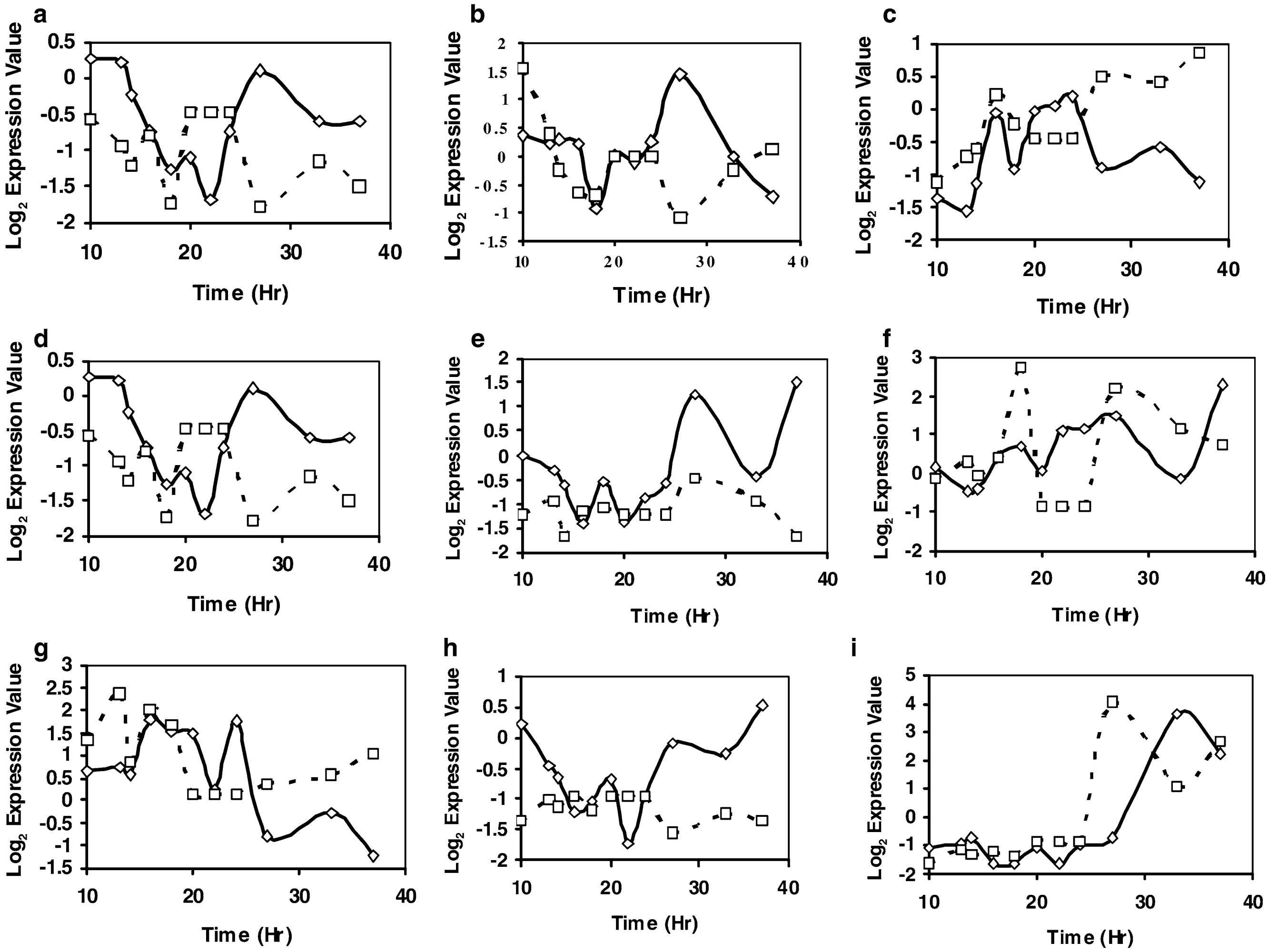

Fig. 9 Expression profile of a representative set of regulatory genes identified as differentially expressed in Δ*absA1* as compared to wild-type *S. coelicolor*. Solid line corresponds to the profile in wild type whereas dotted line corresponds to Δ*absA1*. Annotation of each gene is as follows: SCO2426: possible regulator protein (**a**); SCO1926: putative DNA-binding protein (**b**); SCO5872 (kdpE):putative turgor pressure two-component regulator (**c**); SCO0940: putative marR-family regulator protein (**d**); SCO1616: putative transcriptional regulator (**e**); SCO7785: putative transcriptional regulator (**f**); SCO0253: putative tetR-family regulator protein (**g**); SCO3736: putative RNA polymerase ECF sigma factor (**h**); SCO5584 (glnB) nitrogen regulatory protein (**i**). Each of these genes belongs to a different cluster when classified using a K-means clustering algorithm. The clustering was based on a concatenated vector of expression profile of each gene in the wild type and Δ*absA1*

genes with low mean and high variance were excluded from consideration. The cut-off Euclidean distance value for identifying differentially expressed genes was chosen based on the mean and the standard deviation of the Euclidean distance over the entire genome. For a Euclidean distance threshold of 3 (mean + 1.2×standard deviation), 491 genes were identified as kinetically differentially expressed. These genes are potentially directly or indirectly controlled by the disrupted gene in the mutant. Note that the genes are classified as differentially expressed based on the entire time profile, and not merely from a single time point. Hence such genes are referred to as kinetically differentially expressed, as distinct from differential expression at a single time point. As shown in Fig. 8, these kinetically differentially expressed genes are distributed over a wide range of functional classes. Among these, more than 70 are involved in secondary metabolism. Detailed physiological analysis of these genes will be presented elsewhere (manuscript under preparation). To elucidate the transcriptional network, genes potentially playing a regulatory role are of special interest. These include genes from four major classes: defined family regulators, sigma and anti-sigma factors, two-component systems, and other genes classified as transcriptional regulators. Out of the 491 differentially expressed genes, 40 genes belonged to these classes with 14, 4, 7, and 15 in each class, respectively.

To discern the effect of *absA1* knockout on the dynamics of the differentially expressed genes, we further classified the kinetically differentially expressed genes based on the correlation coefficient between the

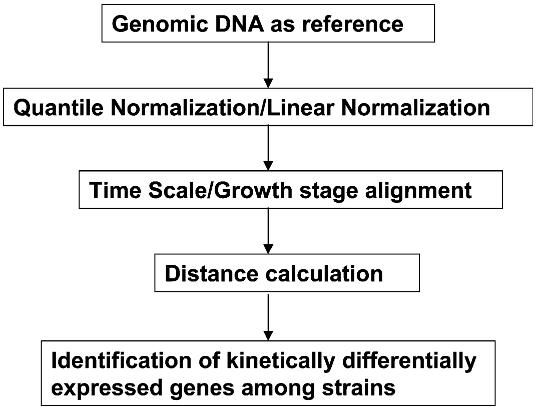
Genomic DNA as reference
Quantile Normalization/Linear Normalization
Time Scale/Growth stage alignment
Distance calculation
Identification of kinetically differentially expressed genes among strains

Fig. 10 Flowchart of overall strategy to compare transcriptional profiles of two strains to identify differentially expressed genes

gene expression profile of wild type and Δ*absA1*. The first class consists of 81 genes, where the correlation coefficient is higher than a chosen threshold of 0.7. These genes exhibit similar profiles in both strains, with only a change in the magnitude of expression. The second class consists of genes where the dynamic profile is perturbed in the knockout strain. To capture the varying effects on the dynamics, the genes were clustered using the k-means clustering algorithm in Spotfire. The time profile of each gene in wild type and Δ*absA1* was concatenated to form a vector, and subsequently was used to classify the genes into 30 clusters. A few example regulatory genes from different clusters are shown in Fig. 9, illustrating the varying effects of Δ*absA1* knockout on the dynamics of these genes. As data accumulate from different strains and culture conditions, we also aim to identify gene sets that bear characteristics of each stage in the growth. These growth marker genes can be subsequently used to align cultures. Identification of such gene clusters can also lead to identification of local clocks that may be decoupled from the global clock.

Accumulation of time series data from the studies of additional mutants will facilitate the identification of causality relationships between regulatory genes ultimately resulting in the building of transcriptional regulatory networks. Various methods have been proposed to this end such as time-lagged correlation analysis, which finds pairs of genes that have similar profiles with one lagging the other [19]; [10]. Various computational frameworks based on Boolean networks [16], dynamic Bayesian networks [23, 24], and Hidden Markov models [18] have been proposed to model time series and gene perturbation data to build the complete gene network.

Concluding Remarks

The genome of *S. coelicolor* encodes more than 7,000 genes that display vastly dynamic behavior during cultivation in liquid medium. This large genetic reservoir and high dynamism of transcriptome probably reflect the microorganism's need to respond to drastic changes in the native habitat. In spite of much progress in the past two decades, the regulatory networks of *S. coelicolor* are still fragmentary. Many genetic elements are yet to be discovered. As in any microarray assay, the identification of differentially expressed transcripts may lead to statistically confident outcomes in some cases. In others, microarray results provide hints as to potential targets for further investigation. For investigating the regulatory gene network for which many genetic elements involved are yet to be identified and pinpointing the targets for further investigation, microarray analysis is a valuable tool.

The transcriptional dynamism and the complex behavior of *S. coelicolor* makes it an intriguing subject for applying DNA microarray analysis for discovery of potential regulatory elements and elucidating the regulation of genetic networks. Key to the discovery is the identification of kinetically differentially expressed transcripts amongst different strains. Our approach represents a systematic way of comparing time series transcription profiles between a disruption mutant and wild-type strain of *S. coelicolor*. The methodology, summarized in Fig. 10, is generally applicable to a larger number of strains.

Such systematic exploration of the genetic and physiological space can potentially lead to a road map for the metabolic engineering of antibiotic production in these microorganisms.

Acknowledgements This work was supported in part by a grant from the National Institutes of Health GM55850. The bioinformatics support was provided by University of Minnesota Supercomputing Institute. SM was supported by a Graduate School Fellowship from the University of Minnesota.

References

1. Aach J, Church GM (2001) Aligning gene expression time series with time warping algorithms. Bioinformatics 17:495–508
2. Baltz RH (1998) Genetic manipulation of antibiotic-producing *Streptomyces*. Trends Microbiol 6:76–83
3. Bar-Joseph Z (2004) Analyzing time series gene expression data. Bioinformatics 20:2493–503
4. Bentley SD, Chater KF, Cerdeno-Tarraga AM, Challis GL, Thomson NR et al (2002) Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature 417:141–147
5. Bibb M (1996) 1995 Colworth prize lecture. The regulation of antibiotic production in *Streptomyces coelicolor* A3(2). Microbiology 142 (Pt 6): 1335–1344
6. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19:185–193
7. Brana AF, Wolfe S, Demain AL (1986) Relationship between nitrogen assimilation and cephalosporin synthesis in *Streptomyces clavuligerus*. Arch Microbiol 146:46–51
8. Bystrykh LV, Fernandez-Moreno MA, Herrema JK, Malpartida F, Hopwood DA, Dijkhuizen L (1996) Production of actinorhodin-related "blue pigments" by *Streptomyces coelicolor* A3(2). J Bacteriol 178:2238–2244

9. Chater KF, Bibb MJ (1997) Regulation of bacterial antibiotic production. In: Rehm H-J, Reed G (eds) Products of Secondary Metabolism (Biotechnology, vol 7), Weinheim: VCH, pp 57–105
10. Filkov V, Skiena S, Zhi J (2002) Analysis techniques for microarray time-series data. J Comput Biol 9:317–330
11. Gadgil M, Lian W, Gadgil C, Kapur V, Hu WS (2005) An analysis of the use of genomic DNA as a universal reference in two channel DNA microarrays. BMC Genomics 6: 66
12. Hopwood DA, Chater KF, Bibb MJ (1995) Genetics of antibiotic production in *Streptomyces coelicolor* A3(2), a model streptomycete. Biotechnology 28:65–102
13. Huang J, Lih CJ, Pan KH, Cohen SN (2001) Global analysis of growth phase responsive gene expression and regulation of antibiotic biosynthetic pathways in *Streptomyces coelicolor* using DNA microarrays. Genes Dev 15:3183–3192
14. Kerr MK, Churchill GA (2001) Experimental design for gene expression microarrays. Biostatistics 2:183–201
15. Kieser T, Bibb MJ, Buttner MJ, Chater KF, Hopwood DA (2000) Practical *Streptomyces* Genetics. The John Innes Foundation, Norwich:ISBN 0-7084-0623-8
16. Mehra S, Hu WS, Karypis G (2004) A Boolean algorithm for reconstructing the structure of regulatory networks. Metab Eng 6:326–339
17. Sankoff D, Kruskal JB (1983) Time warps, string edits, and macromolecules : the theory and practice of sequence comparison Advanced Book Program xii. Addison-Wesley, Reading 382 pp
18. Schliep A, Schonhuth A, Steinhoff C (2003) Using hidden Markov models to analyze gene expression time course data. Bioinformatics 19 (Suppl 1):i255–i263
19. Schmitt WA Jr, Raab RM, Stephanopoulos G (2004) Elucidation of gene interaction networks through time-lagged correlation analysis of transcriptional data. Genome Res 14:1654–1663
20. Talaat AM, Howard ST, Hale Wt, Lyons R, Garner H, Johnston SA (2002) Genomic DNA standards for gene expression profiling in Mycobacterium tuberculosis. Nucleic Acids Res 30:e104
21. Williams BA, Gwirtz RM, Wold BJ (2004) Genomic DNA as a cohybridization standard for mammalian microarray measurements. Nucleic Acids Res 32:e81
22. Yang YH, Dudoit S, Luu P, Lin DM, Peng V et al (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res 30:e15
23. Yu J, Smith VA, Wang PP, Hartemink AJ, Jarvis ED (2004) Advances to Bayesian network inference for generating causal networks from observational biological data. Bioinformatics 20:3594–3603
24. Zou M, Conzen SD (2005) A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics 21:71–9