Metabolic modeling of Streptomyces and its relatives

Alam, Mohammad Tauqeer

Metabolic modeling of *Streptomyces* and its relatives:

A constraints-based approach

The work described in this thesis was carried out at the Groningen Bioinformatics Centre, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, The Netherlands, and at The Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, Joseph Black Building, University of Glasgow, United Kingdom. The research was financially supported by Groningen Biomolecular Sciences and Biotechnology Institute (GBB), University of Groningen, The Netherlands.

Thesis front cover: A minimal enzyme-association network of *Streptomyces clavuligerus*, constructed from a genome-scale metabolic model by linking enzyme nodes via edges based on shared common metabolites. *In silico* predictions of single enzyme knockouts, and conditional enzyme knockouts are superimposed on the network, and enzyme essentiality is indicated by node color: red and orange nodes represent "unconditional" and "conditional" essential enzymes respectively. White nodes represent enzymes which connects essential enzymes with conditionally essential enzymes ("gaps"), and green nodes are non-essential enzymes which can be removed from the model to minimize the genome.

Thesis back cover: A *Streptomyces* colony. Courtesy Prof. Dr. E. Takano.

Printed by: Ipskamp Drukkers Nijmegen.

ISBN: 978-90-367-4951-0 (book)
ISBN: 978-90-367-4952-7 (digital)

RIJKSUNIVERSITEIT GRONINGEN

Metabolic modeling of *Streptomyces* and its relatives:

A constraints-based approach

Proefschrift

ter verkrijging van het doctoraat in de
Wiskunde en Natuurwetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. E. Sterken,
in het openbaar te verdedigen op
vridag 24 juni 2011
om 16.15 uur

door

Mohammad Tauqeer Alam

geboren op 9 januari 1981
te Madhubani, Bihar, India

Promotores	: Prof. dr. R. Breitling Prof. dr. R. C. Jansen
Beoordelingscommissie	: Prof. dr. M. Barrett Prof. dr. B. Teusink Prof. dr. M. Heinemann

بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ

To my parents.

Contents

Abstract

Streptomyces species are often referred to as "antibiotic factories" due to their ability to produce a large number of clinically important compounds. They belong to the order *Actinomycetales*, which is biologically very diverse, showing differences in genome size, pathogenicity, ecological niche, as well as in the ability of some of the species to produce various secondary metabolites.

This thesis starts by introducing the genus *Streptomyces* and its relatives and describing the modeling techniques used for analyzing their metabolic functions (Chapter 1). We explored the metabolic system of two antibiotic producing model bacteria, *Streptomyces coelicolor* (Chapter 2) and *Streptomyces clavuligerus* (Chapter 3), and computationally investigated their mechanism of antibiotic production. To understand how these antibiotic producing species are phylogenetically related to other species of the group *Actinomycetales*, we constructed a comprehensive phylogenetic tree, and established a generally usable robust approach to construct fully resolved phylogenetic trees from genome sequences (Chapter 4). Results of the phylogenetic study formed the basis for large-scale metabolic modeling, and we identified metabolic as well as topological commonalities and differences among members of the group (Chapter 5). Furthermore, by combining phylogenetic information with gene expression data we prioritized "orphan" genes of *Streptomyces coelicolor* for future experimental study (Chapter 6).

Finally, the thesis concludes by discussing the future use of our results and models and outlines some perspective for further research into the Systems biology of antibiotic producing microbes (Chapter 7).

Chapter 1

Introduction

In the past 15 years, after the genome of first cellular organism *H. influenzae* was sequenced (Fleischmann et al., 1995), the availability of metabolic network models has rapidly increased. Now, with the advent of high-throughput next-generation sequencing technologies, the number of sequenced genomes are exponentially increasing, and it is now possible to sequence genomes of rarely studied organisms, or entire groups of organisms. Similar rapid development of high-throughput technologies has occurred in other 'omics' fields, including transcriptomics, proteomics and metabolomics. We can now comprehensively characterize cellular systems at different levels. However, the main challenge is to utilize this vast amount of high-throughput biological data to understand and predict the function of the studied system as a whole.

In this regard, several different computational approaches have been developed and used extensively in the field of Systems Biology (Gombert and Nielsen, 2000; Arkin, 2001; Varma and Palsson, 1994). One powerful approach is constraint-based flux balance analysis (Varma and Palsson, 1994), which takes the entire metabolic system of a cell into consideration and predicts the profile of metabolic fluxes of a complete organism *in silico* (Bonarius et al., 1997; Covert et al., 2001a; Price et al., 2004; Marco et al., 2009). Several studies integrating different biological data with these constraints-based models have been successfully performed recently (Covert et al., 2001b; Covert and Palsson, 2002; Akesson et al., 2004; Feist and Palsson, 2008; Oberhardt et al., 2009; Fleming et al., 2010; Alam et al., 2010b).

This thesis focuses on the construction and application of genome-scale models of two bacteria of the genus *Streptomyces*, model organisms of a large group of biotechnologically important organisms. These two separate studies were followed by constructing a robust and fully resolved phylogenetic tree of *Streptomyces* and its relatives in the group *Actinomycetales* which formed the basis for a large comparative modeling study of all genome-sequenced strains of *Streptomyces* and their relatives.

The following sections will introduce *Streptomyces* and its relatives, describe the modeling techniques used for analyzing the metabolic system of these organisms and present a brief outline of the structure of this thesis.

1.1 *Streptomyces* and its relatives

Streptomyces bacteria are known for their ability to produce a vast diversity of secondary metabolites, including some very important commercially available antibiotics (Hopwood, 2007). These organisms belong to one of the best studied and most diverse orders of bacterial taxonomy, *Actinomycetales*, of the phylum *Actinobacteria* (Lechevalier and Lechevalier, 1967; Embley and Stackebrandt, 1994; Hopwood, 2007; Ventura et al., 2007). Actinomycetes are gram positive bacteria with a very high genomic G+C content. They exhibit vast biodiversity in terms of their genome size, ability to survive in extremely different environmental conditions, the pathogenicity of some of the organisms, as well as the varying capability to synthesize specific secondary metabolites (Lechevalier and Lechevalier, 1967; Embley and Stackebrandt, 1994; Hopwood, 2007; Ventura et al., 2007). Despite their remarkable diversity, these organisms are monophyletic, i.e. they are located in one branch of the phylogenetic tree (Alam et al., 2010a). This makes it easier to understand how these organisms have evolved, and to investigate how they manage to thrive in their specific environment aided by their distinct characteristics.

We have used genome sequences to build a comprehensive, detailed and robust phylogenetic tree of these organisms by combining several single-sequence based approaches with some whole-genome based approaches to elucidate the precise phylogenetic relationships within *Actinomycetales*. This was used as a reference for a large comprehensive metabolic modeling study as described below. The individual phylogenetic analysis approaches have been described in detail in chapter 4.

1.2 Reconstruction and analysis of genome-scale metabolic networks

The genomes of thousands of microorganisms have been sequenced, including those of several actinomycetes. This information can be used to reconstruct the metabolic network of these organisms on a genomic scale. In the past, when there were no annotated genome sequences, the metabolic networks of a few well-studied model organisms were analyzed by collecting relevant information from extensive body of literature and through detailed biochemical characterization of purified enzymes. Example of such studies are the analysis of, acetate over-flow in *E. coli* (Majewski and Domach, 1990), the stoichiometric interpretation of glucose catabolism in *E. coli* (Varma et al., 1993) and the production of oxychemical in *B. subtilis* (Papoutsakis and Meyer, 1985). Now, with the availability of genome sequences, we can generate metabolic models at a genome-wide scale, even for less-well studies organisms or an entire group of organisms and perform comparative modeling.

The detailed process of reconstructing a genome-scale metabolic model has been described and reviewed elsewhere (Feist et al., 2009; Thiele and Palsson, 2010). Briefly, the fundamental steps to generate a genome-scale metabolic model are: automated generation of preliminary genome-based reconstruction, curation of the preliminary reconstruction, transforming the curated models into a mathematical model, and integration of high-throughput data to improve the reconstruction (Feist et al., 2009; Thiele and Palsson, 2010). Genome annotation is used to generate a draft reconstruction, which yields an initial set of biochemical reactions encoded in a genome; however these automatically generated models contain gaps, as well as some wrong annotations. There are several efficient algorithms which can detect and fill gaps automatically (Kumar et al., 2007; Henry et al., 2010); however one should carefully check these fillings manually (Feist et al., 2009; Thiele and Palsson, 2010). Once the gaps are filled, the next step is to translate the reconstruction into a computational model which can be used to check its physiological capabilities in different biological environments. To refine the mathematical model, different sets of high-throughput data are used. In this step, an iterative process is often used to identify discrepancies between model predictions and observed phenotypes.

Possible applications of genome-scale models are enormously diverse (Feist and Palsson, 2008; Oberhardt et al., 2009). These models can be used in the context of several high-throughput data, for instance, gene expression data, protein expression data, or metabolomics profiles. Based on experimental observations, if a certain pathway is significantly active, the corresponding flux can be constrained to fall within a specified range or have a specific value (Akesson et al., 2004; Alam et al., 2010b; Medema et al., 2011b, 2010). Similarly, gene expression data can be used to constrain the set of available reactions (Akesson et al., 2004). These experimentally observed data can be superimposed on a network to investigate significantly changed pathways or "metabolic hotspots". Metabolic engineering can also be applied to these models by selectively altering the metabolism of the cell and to investigate the expected consequences of planned interactions. Due to the availability of thousands of genome sequences and powerful pipelines to generate analysis-ready reconstructions, it is now possible to perform a comparative modeling to understand the metabolic relation of large sets of species and to investigate the metabolic commonalities and differences among groups of organisms.

1.3 Constraint-based flux balance analysis

After the completion of an analysis-ready genome-scale metabolic model, the metabolic reactions are mathematically represented in the form of a stoichiometric matrix (S) of size $m \times n$, where m (the number of rows) represents unique metabolites, n (the number of columns) represents different reactions of the model, and entries of the matrix indicate the stoichiometry of metabolites participating in the reactions. A substrate which is consumed in a reaction has a negative coefficient whereas products of the reaction have a positive coefficient. The stoichiometric matrix (S) is a sparse matrix, because only a few metabolites are involved in most biochemical reactions. v is the vector of all reaction fluxes of the network (Palsson, 2006).

At steady state $S * v = 0$, which corresponds to a set of linear equations. Since the number of reactions in any metabolic model is usually larger than the number of metabolites ($n > m$), or in other words, the number of unknowns is larger than the number of variables, there is no unique solution to this set of equations. However, by imposing different constraints on the system, and optimizing a particular objective function, e.g. maximize the flux through a specific reaction or set of reactions, we can find a single point of the solution space (Palsson, 2006). Typical objective functions are: maximizing growth, maximizing ATP production, minimizing nutrients consumption or maximizing antibiotics production.

The two main class of constraints are mass-balance constraints and flux bounds. Mass-balance constraints ensure that the amount of metabolites consumed in the systems must equal the amount of metabolites produced. Flux bounds are imposed by fixing minimum and maximum values for each reaction flux, based on reaction reversibility, environmental conditions and sometimes experimental information. The model construction approach and the main concepts of flux balance analysis are summarized in Figure 1.1.

Flux balance analysis is a very powerful approach to study large and complex metabolic systems, as it can be computed very quickly even for a very large network, and because it does not need any enzyme kinetic parameters. One can perform a large number of *in silico* experimental perturbations rapidly and test many different hypotheses, but the approach also has some limitations; FBA cannot predict metabolite concentrations, it is only suitable for predicting fluxes at steady state, and it does not account for regulatory effects in its standard implementation, therefore, its prediction always not be accurate (Famili et al., 2003).

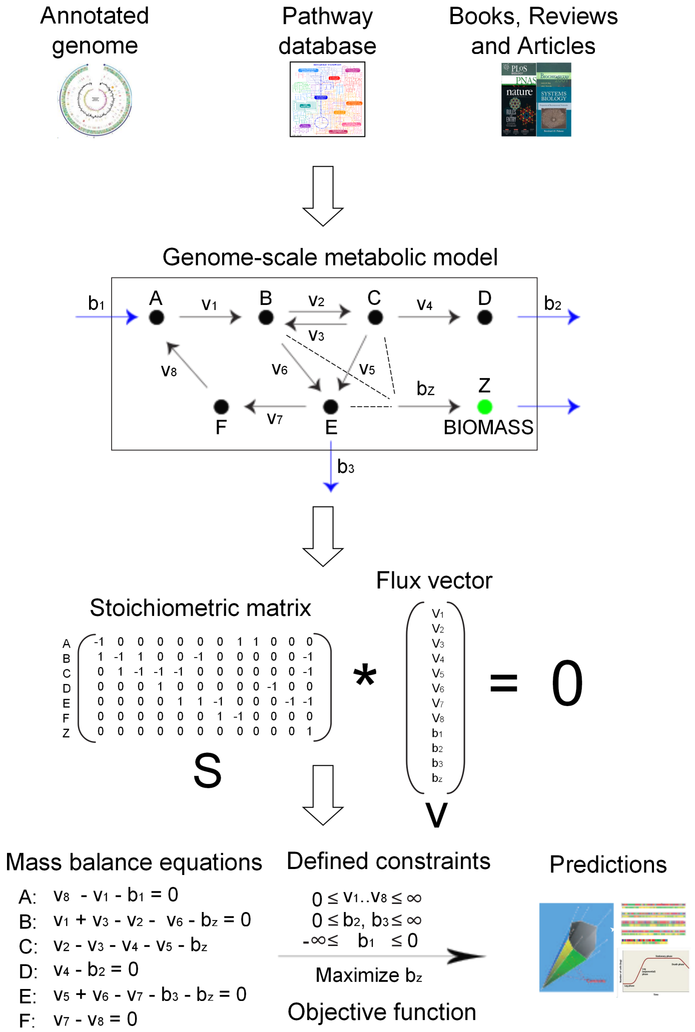
Annotated genome
Pathway database
Books, Reviews and Articles
Genome-scale metabolic model
b1
A
v1
B
v2
v3
C
v4
D
b2
v8
v6
v5
bz
Z
F
v7
E
BIOMASS
b3
Stoichiometric matrix
A -1 0 0 0 0 0 0 1 1 0 0 0
B 1 -1 1 0 0 -1 0 0 0 0 0 -1
C 0 1 -1 -1 -1 0 0 0 0 0 0 -1
D 0 0 0 1 0 0 0 0 0 -1 0 0
E 0 0 0 0 1 1 -1 0 0 0 -1 -1
F 0 0 0 0 0 0 1 -1 0 0 0 0
Z 0 0 0 0 0 0 0 0 0 0 0 1
S
*
Flux vector
v1 v2 v3 v4 v5 v6 v7 v8 b1 b2 b3 bz
V
= 0
Mass balance equations
A: v8 - v1 - b1 = 0
B: v1 + v3 - v2 - v6 - bz = 0
C: v2 - v3 - v4 - v5 - bz
D: v4 - b2 = 0
E: v5 + v6 - v7 - b3 - bz = 0
F: v7 - v8 = 0
Defined constraints
0 ≤ v1..v8 ≤ ∞
0 ≤ b2, b3 ≤ ∞
-∞ ≤ b1 ≤ 0
Maximize bz
Objective function
Predictions

Figure 1.1: ***General outline of the reconstruction of genome-scale metabolic models and principles of flux balance analysis.*** *Information from different sources, including annotated genome sequences, pathway databases, and biochemistry books and reviews, are collected together for the reconstruction of a genome-scale model. A hypothetical reaction for biomass formation and a specific set of exchange reactions are defined from books and literatures and used in FBA for optimization and constraining the model, respectively. The example network contains*

6 metabolites (A – F) and 12 reactions, including 8 metabolic reactions, 3 exchange reaction and 1 hypothetical reaction of biomass formation. The network is mathematically represented in the form of stoichiometric matrix, S, *where 7 rows represent compounds, including 6 metabolites (A–F) and a biomass* Z, *and 12 columns represent reactions. The vector* v *is a column vector of 12 flux variables. At steady state,* S * v = *0, which gives a set of linear mass balance equations. By imposing a set of constraints the solution space can be reduced. One optimal solution of the system of equation is achieved by optimizing an objective function (biomass production in this case) and the obtained optimal flux profile is used for predictions.*

1.4 Thesis contribution and organization

In this thesis, I present an analysis of genome-scale metabolic models of several actinomycetes, along with a detailed genome-based phylogenetic analysis of the entire group.

Chapter 2 describes a computational study of exploring the metabolic switch from primary phase to the secondary phase during the growth of *Streptomyces coelicolor*. In the primary phase of growth, cells grow exponentially; when the available nutrients get depleted they then switch to stationary phase with a major reorganization of metabolism, in particular the activate production of secondary metabolites. For this analysis we constructed a genome-scale metabolic model of *Streptomyces coelicolor* and incorporated additional qualitative constraints based on online measurements from a large cultivation experiment; we then integrated model predictions and detailed gene expression data to understand the mechanisms involved in metabolic switch.

In Chapter 3, we present the genome-scale metabolic model of a second antibiotic producing organism, *Streptomyces clavuligerus*. Here, we studied genome-wide changes in gene expression of an antibiotic high producer strain, obtained by a random mutagenesis and selection approach, with that of a parental wild type strain. We performed flux balance analysis and optimized biomass and antibiotic production to identify key upregulated reactions. The study provided new insight into the metabolic changes accompanying the mutations that had led to antibiotic overproduction.

As a prerequisite for a large comparative modeling, we analyzed the phylogeny of entire group of organisms, which is presented in **Chapter 4**. To make a single reliable complete resolved consensus tree of the group, we used information from several different levels, including individual gene sequences as well as a multitude of whole-genome based phylogenetic approaches. We were able to obtain a fully resolved consensus tree and at the same time established a generally usable strategy for the robust phylogenetic analysis of large classes of genome-sequenced organisms.

In the research described in **Chapter 5**, we used the phylogenetic results to identify metabolic trends among actinomycetes on a systems-scale. We constructed genome-scale metabolic models for the entire group of organisms and analyzed their metabolism in one common minimal growth medium. We performed single *in silico* gene knockouts to identify broader patterns of gene essentiality. We also discuss general topological features of the metabolic networks of this diverse class of organisms as revealed by the genome-scale comparative modeling.

Finally in **chapter 6**, we used phylogenetic analysis of Chapter 4 and the gene expression of Chapter 2 to prioritize ”orphan” genes for future experimental study.

Chapter 7 of this thesis contains concluding remarks and describes future perspectives for the field.

Published as: **Alam MT**, Merlo ME, The STREAM Consortium, Hodgson DA, Wellington EMH, Takano E, Breitling R (2010) : *Metabolic modeling and analysis of the metabolic switch in Streptomyces coelicolor* **BMC Genomics 2010, 11:202**

Chapter 2

Metabolic modeling and analysis of the metabolic switch in *Streptomyces coelicolor*

Abstract

Background: *The transition from exponential to stationary phase in* Streptomyces coelicolor *is accompanied by a major metabolic switch and results in a strong activation of secondary metabolism. Here we have explored the underlying reorganization of the metabolome by combining computational predictions based on constraint-based modeling and detailed transcriptomics time course observations.*

Results: *We reconstructed the stoichiometric matrix of* S. coelicolor, *including the major antibiotic biosynthesis pathways, and performed flux balance analysis to predict flux changes that occur when the cell switches from biomass to antibiotic production. We defined the model input based on observed fermenter culture data and used a dynamically varying objective function to represent the metabolic switch. The predicted fluxes of many genes show highly significant correlation to the time series of the corresponding gene expression data. Individual mispredictions identify novel links between antibiotic production and primary metabolism.*

Conclusion: *Our results show the usefulness of constraint-based modeling for providing a detailed interpretation of time course gene expression data.*

2.1 Background

The transition from exponential growth to stationary phase is a major event in microbial physiology (Kolter et al., 1993). During the exponential phase of growth, bacterial cells produce metabolites necessary for growth and grow rapidly. Once essential nutrients have been depleted, cells switch to stationary phase, stop growing, reorganize their energy metabolism and often start producing a new set of secondary metabolites, including antibiotics (Roszak and Colwell, 1987). In this study, we have explored the metabolic switch in *Streptomyces coelicolor*, the model organism of the antibiotics producing genus *Streptomyces*. The genome of this soil bacterium has been sequenced and contains about 7825 genes, one of the largest numbers for any bacterium (Bentley et al., 2002). More than 20 clusters coding for the 4 known and several predicted antibiotics or related compounds have been identified in the genome (Challis and Hopwood, 2003). To optimize the production of valuable secondary metabolites, understanding the shift from primary to secondary metabolism during the transition phase will play a key role.

We constructed a constraints-based genome-scale stoichiometric model of *S. coelicolor* metabolism, based on earlier similar models (Borodina et al., 2005a, 2008), and integrated the model predictions with a large gene expression dataset (Nieselt et al., 2010). The constraints-based approach, in particular flux balance analysis, has been shown to be highly predictive of growth phenotypes in many microbial systems (Price et al., 2003, 2004) and can be used to construct large scale metabolic models based on genome sequences in the absence of kinetic information, making it particularly attractive for less well-studied organisms like *S. coelicolor*.

Predictions from constraint-based models usually hold for steady-state assumptions (Durot et al., 2009; Lee et al., 2006). To enable the incorporation of experimental information from timeseries measurements, we extend the approach by applying a dynamically changing input function (specifying constrains on nutrient uptake) and objective function (specifying the shift of cellular resources from cellular growth to antibiotics production). The predicted flux profiles are then compared to the gene expression profiles of the corresponding enzyme-coding genes to validate the model.

We observe a surprisingly good correlation between predicted fluxes and measured gene expression, indicating both the power of the constraint-based modeling approach and the tight regulation of gene expression in *S. coelicolor*. A small number of incorrectly predicted fluxes indicate the need for including additional gene regulatory constraints to the model (Covert et al., 2001b; Lee et al., 2008), but also allows the sensitive identification of misannotations and putative novel reactions involved in secondary metabolite biosynthesis.

2.2 Results and Discussion

We have reconstructed a genome-scale model of *Streptomyces coelicolor* metabolism with recent updated annotations as discussed in the Methods section. Our aim was to study the metabolic switch between the primary phase and secondary phase of growth.

2.2.1 Initial model validation

To validate our model we first compared predicted growth rates to those reported for glucose limited environments (Melzoch et al., 1997). In that work, *S. coelicolor* had been grown in chemostat culture in a chemically defined medium under various nutrient limitations. As the dilution rate of the chemostat is the same as the specific growth rate at steady state we can compare it directly to the prediction of the *in silico* model. In our model we used an input function that mimics the glucose limited medium used in these experiments, adopting the observed glucose and oxygen uptake rate as well as carbon dioxide and actinorhodin production rates as initial conditions in the model. We maximized biomass production to predict the optimized *in silico* specific growth rate and we compared the predicted growth with the observed growth. Figure 2.1 and Table 2.1 show that observation and prediction are in good agreement, indicating the general validity of our model.

2.2.2 Global metabolic switching from primary phase to secondary phase of growth

For a more detailed understanding of the metabolic transition phase, we then modeled flux changes happening during fermentation culture on phosphate

Table 2.1: **Comparison of experimentally observed dilution rates from chemostat data Melzoch et al. (1997) and predicted specific growth rates**

Glucose (mmol/g.h)	O2 (mmol/g.h)	CO2 (mmol/g.h)	Actinorhodin (μg/g.h)	Observed dilution rate D (/h)	Predicted specific growth rate μ(/h)
0.5	1.8	1.9	2	0.035	0.0272
0.6	2	2	2	0.045	0.0396
0.8	2.4	2.5	415	0.06	0.0539
0.9	2.5	2.7	152	0.072	0.0657
1.1	3.1	3.1	60	0.092	0.0862
1.85	6.6	6.7	7	0.115	0.1088
2.1	7.2	7	5	0.128	0.1385

limiting medium. For this growth condition we had earlier collected a detailed gene expression time series. Based on the measured nutrient uptake and product formation, we dynamically adapted the objective function and optimized the *in silico* specific growth rate. The optimum specific growth rate and optimal flux vector for all metabolic reactions were predicted for each time point. Figure 2.2 shows the observed normalized depletion of substrate glucose, glutamate and phosphate during growth. At about 34 hour, phosphate is depleted, triggering the transition to stationary phase and the production of antibiotics by the bacteria. The corresponding slow-down of growth matches well between prediction and observation.

Next we compared the predicted metabolic flux profile of all 549 enzyme-coding genes to the corresponding gene expression data from Nieselt et al. (Nieselt et al., 2010). A histogram of correlation coefficients between predicted flux and observed gene expression is shown in figure 2.3. The correlation of predicted flux and gene expression level is highly significant, and a large number of genes exhibit very high correlation (33% of genes; $r > 0.5$). This shows not only the global validity that these genes of our model are probably correctly annotated in the model but also illustrates the tight regulation of gene expression level for enzyme-coding genes of *S. coelicolor*.

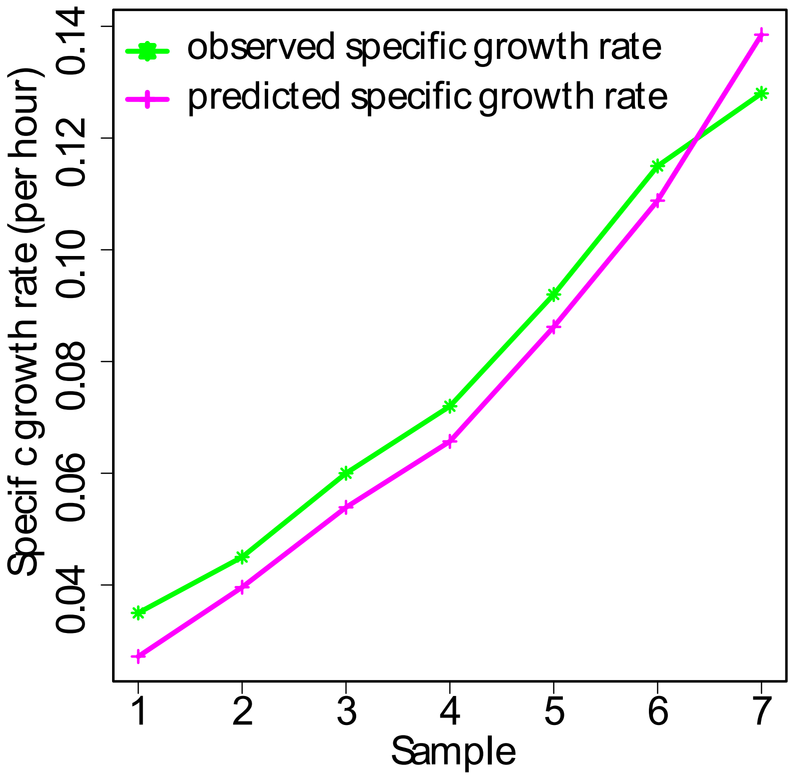

Figure 2.1: ***Model validation.*** *Comparison of the experimentally observed specific growth rate from chemostat data (Melzoch et al., 1997) and the predicted* in silico *specific growth rate from the model in glucose limited media. The specific rate of glucose consumption, oxygen consumption, carbon dioxide production and actinorhodin production from 7 different conditions were taken from (Melzoch et al., 1997) and used as initial condition in the model.*

This is in agreement with the general observation that gene expression is more tightly regulated in unicellular compared to multicellular organisms, for evolutionary reasons, such as the much larger effective population size and stronger energetic constraints in small organisms (Wagner, 2005, 2007).

A large set of genes does not show correlation (64% of genes; –0.5 < r < 0.5). These are mostly genes that do not change expression (nor predicted flux) along the time course. In these cases of constant expression no cor-

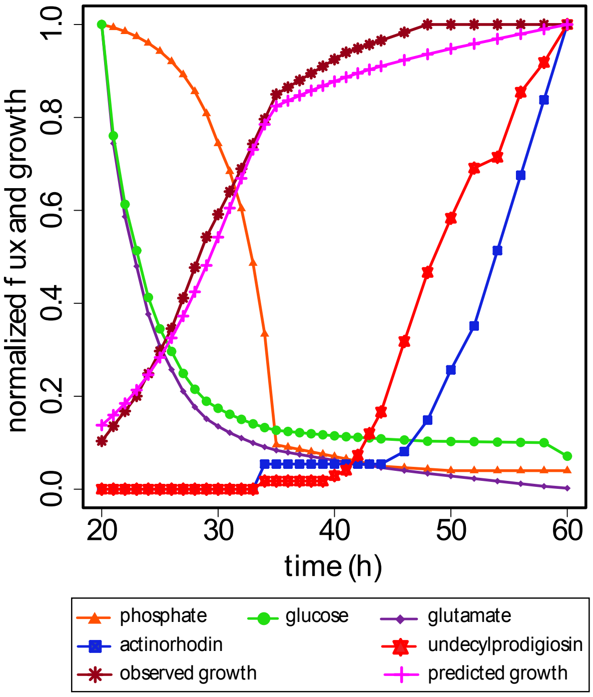

Figure 2.2: ***Dynamic model constraints and predicted cell growth.*** *Based on online measurement on a fermenter experiment, normalized constraints of model influx of phosphate, glucose, and glutamate and the production of the antibiotics actinorhodin and undecylprodigiosin were determined. Their time course is shown together with the experimentally observed and* in silico *predicted growth.*

relation information is present in the data, leading to correlation coefficient close to zero. Of course, there will also be cases where gene expression levels and flux levels do not correlate for other reasons, for instance due to post-transcriptional and post-translational regulation mechanisms.

Strikingly, there is also a small group of strongly anticorrelating genes (15 genes; $r < -0.5$). These are potentially the most interesting cases; they could indicate wrong annotations of gene function, but also the unexpected presence of regulatory constraints or novel functionalities of genes. To further examine these options, we subdivided all 549 observed expression profile into 12 clusters, based on unsupervised hierarchical clustering. The number

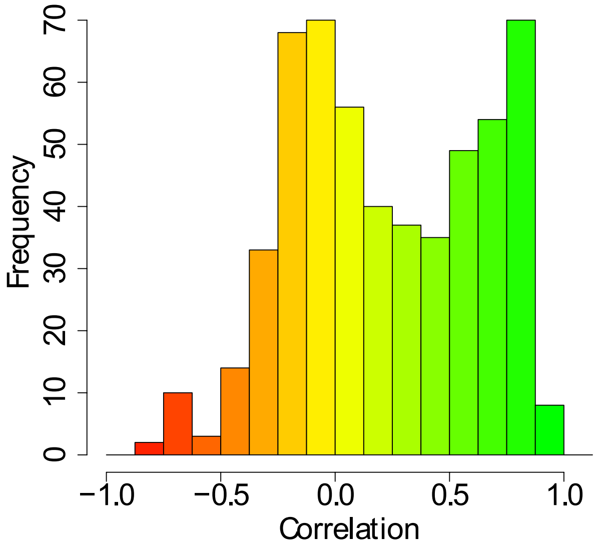

Figure 2.3: ***Correlation between predicted flux and observed gene expression.*** *The histogram shows the correlation between gene expression and predicted flux for 549 enzyme–coding genes. A large number of enzyme–coding genes show high correlation. They include many primary metabolism genes and antibiotic biosynthesis genes. About half of the genes show poor correlation; these are mostly genes that show constant gene expression and/or predicted flux across the entire time course, leading to a correlation coefficient close to zero. A small but noteworthy number of genes show statistically significant negative correlation between gene expression levels and predicted flux. These cases are discussed in more detail in the main text.*

12 was chosen to allow sufficient resolution of different expression pattern. Figure 2.4 shows the average expression time course of each of the 12 resulting clusters. For instance, the pink and red clusters, which contain many genes involved in secondary metabolite production, switch on upon phosphate depletion. The purple, navy blue and blue clusters mostly include genes involved in central metabolism and anabolic functions and are downregulated when nutrient resources in the medium are depleted.

When mapping the gene expression clusters onto the genome (figure 2.5) it is clear that genes with similar expression dynamics tend to be neighbors along the chromosome. Moreover, when also visualizing the correlation be-

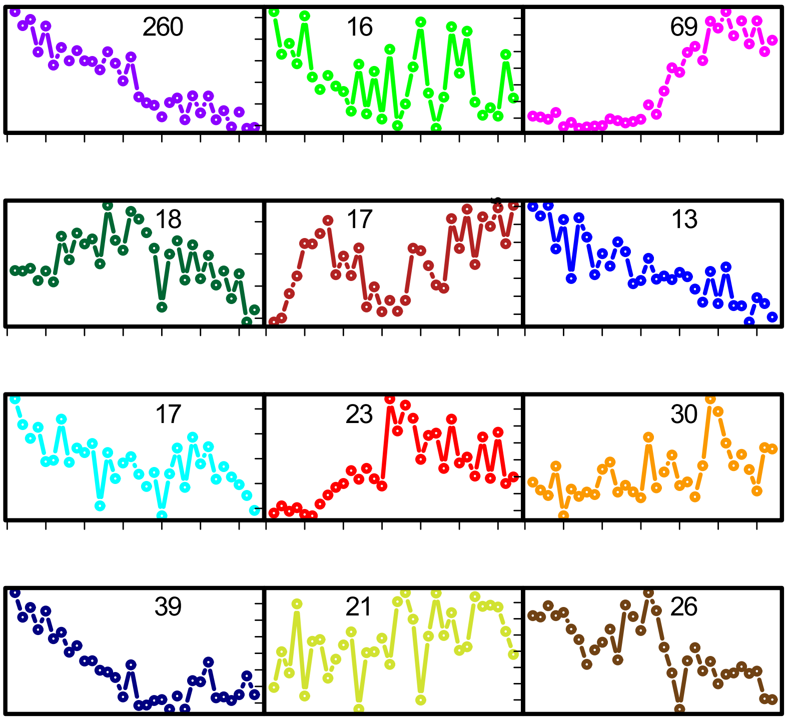

Figure 2.4: ***Average expression profile of 12 expression clusters defined by hierarchical clustering.*** *Gene expression profiles of all enzyme-coding genes in our metabolic flux model were subjected in unsupervised clustering. The number of genes in each cluster is indicated. Several clusters show a clear expression trend matching the changing physiology of the fermentation. The pink cluster is the "antibiotics" cluster, switching on upon phosphate depletion; the purple cluster includes the majority of central metabolism genes that are down-regulated.*

tween gene expression and predicted flux, one can see the that strong anticorrelated expression is seen almost exclusively for genes in the pink and red clusters, which switched on expression during the transition phase, while the predicted flux for these genes are decreasing along the time course (Table 2.2). In contrast, many genes with high positive correlation belong to the purple, navy blue and blue clusters which contain genes of central metabolism, including biosynthesis clusters for arginine, cysteine, glutamate, glutamine, glycine, fatty acid, histidine, homoserine, isoleucine, leucine, lysine, methio-

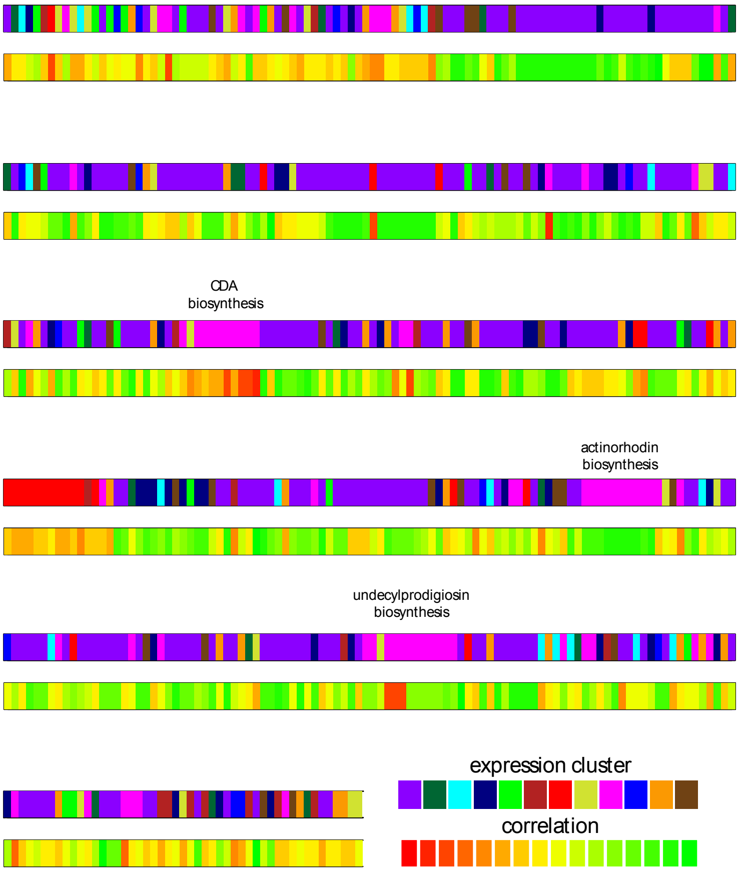

Figure 2.5: ***Genome mapping of expression clusters and correlation between expression and predicted flux.*** *All enzyme-coding genes are shown arranged in their order along the chromosome. The upper trace colors genes according to their membership in one of 12 expression clusters figure 2.4; genes belonging to the same cluster tend to be neighbors along the chromosome, reflecting the operon structure of the genome. The lower trace shows how strongly the predicted flux for each gene correlates with its expression. Genes from some expression clusters tend to show good correlation to the predicted flux (green), e.g. those in the central metabolism cluster (purple); mispredictions (red) seem to cluster along the chromosome and normally affect genes that are upregulated in stationary phase (pink cluster). The position of three major antibiotics biosynthesis clusters is highlighted.*

Table 2.2: **List of anticorrelated genes.** *The most strongly anticorrelated genes are listed on the basis of correlation between gene expression and predicted flux (r<–0.25). Pathway annotations are based on the KEGG database without manual curation. Most of the genes belong to the pink or red cluster which are upregulated in stationary phase; these represent potential new genes involved in antibiotics biosynthesis. Several examples validating this interpretation are discussed in detail in the text.*

SCO ID	Definition	Pathway	r
SCO2286	alkaline phosphatase	folate biosynthesis	–0.80
SCO3249	[acyl-carrier-protein] S malonyltransferase	fatty acid biosynthesis	–0.78
SCO5887	[acyl-carrier-protein] S malonyltransferase	fatty acid biosynthesis	–0.75
SCO0386	asparagine synthetase, glutamine hydrolysing	Aspartate metabolism	–0.75
SCO3248	pentadecanoyl-[acyl-carrier protein] synthesis	fatty acid biosynthesis	–0.74
SCO5886	pentadecanoyl-[acyl-carrier protein] synthesis	fatty acid biosynthesis	–0.71
SCO5888	pentadecanoyl-[acyl-carrier protein] synthesis	fatty acid biosynthesis	–0.70
SCO0828	alkaline phosphatase	folate biosynthesis	–0.67
SCO2068	alkaline phosphatase	folate biosynthesis	–0.66
SCO3246	pentadecanoyl-[acyl-carrier protein] synthesis	fatty acid biosynthesis	–0.63
SCO3595	D-alanine-D-alanine ligase	D-alanine metabolism	–0.63
SCO3221	prephenate dehydrogenase	tryptophan biosynthesis	–0.63
SCO6655	GTP cyclohydrolase II	riboflavin metabolism	–0.61
SCO6787	butyryl-CoA dehydrogenase	propanoate metabolism	–0.59
SCO2687	GTP cyclohydrolase II	riboflavin metabolism	–0.52

nine, N-acetyl muramic acid (NAM) and N-acetyl glucosamine (NAG), as well as sulphate metabolism. Expression of the antibiotic gene clusters for actinorhodin and undecylprodigiosin was also highly correlated to the predicted fluxes.

One large group of anticorrelated genes is the set of 10 genes located the middle of the calcium dependent antibiotics (CDA) biosynthesis gene cluster (SCO3210-SCO3249) (Hojati et al., 2002). SCO3210 and SCO3221 are annotated as 2-dehydro-3-deoxyheptonate aldolase and prephenate dehydrogenase respectively, part of the shikimate pathway (tryptophan biosynthesis). Tryptophan is a precursor for CDA, and there are four anticorrelated genes (SCO3211-3214) which encode for enzymes TrpC2, TrpD2, TrpG, and TrpE2. It seems obvious that these genes are involved in the biosynthesis of tryptophan for CDA biosynthesis and not in the production of tryptophan for general primary metabolism. Indeed it has been shown that these genes do not complement a deficiency in central tryptophan biosynthesis (Hojati et al., 2002). SCO3249 encodes an ACP homolog, and the adjacent genes SCO3246 and SCO3248 along with SCO3228 are proposed to be involved in the biosynthesis of the N-terminal epoxyhexanoyl fatty acid side chain (Hojati et al., 2002). While the direct involvement in CDA biosynthesis has not yet been established for all of these genes, the non-complementation as well as the clear anticorrelation in our model analysis point to the existence of strong regulatory constraints on the expression of these genes. Such regulatory constraints are not routinely included in flux balance analysis, but can substantially enhance its predictive accuracy (Covert et al., 2001b; Lee et al., 2008). Our result shows that a lack of regulatory information can be efficiently compensated by the integration of transcriptomics information, which quite specifically highlights this group of genes for further study.

Another group of anticorrelating genes is seen in the middle of the undecylprodigiosin biosynthesis gene cluster (Challis and Hopwood, 2003; Cerdeo et al., 2001). Three genes (SCO5886, SCO5887 and SCO5888) in this cluster were automatically annotated in our model as fatty acid biosynthesis genes on the basis of sequence similarity with fatty acid genes (3-oxoacyl-[acyl-carrierprotein] synthase II, acyl carrier protein, 3-oxoacyl-[acyl carrier protein] synthase III). However these three genes are well known to be involved in undecylprodigiosin production under the gene names redR (SCO5886), redQ (SCO5887) and redP (SCO5888). This is a clear example of a misannotation that is revealed by the correlation analysis and can easily be fixed in the model.

A third example of strongly anticorrelated genes listed in table 2.2 are three alkaline phosphatases – SCO2286 (phoA), SCO0828 (phoC) and SCO2068 (phoD) – which are assigned in the KEGG database (and consequently in our model) to the folate biosynthesis pathway. Their expression pattern, which shows strong induction upon phosphate depletion, is consistent with earlier reports on their control by PhoR/PhoP (Apel et al., 2007) and a potential role in secondary metabolism, but is less easy to reconcile with a putative function in folate biosynthesis, which is based only on sequence homology.

In all three of these cases, the integration of gene expression and model flux predictions highlighted groups of genes involved in antibiotics production. A small set of additional anticorrelated genes (Table 2.2) are widely scattered through out the genome (Figure 2.5). Each of them is a potential candidate from model correction and for the identification of new secondary metabolite biosynthesis genes with specifically constrained gene expression patterns.

Our biological understanding of *S. coelicolor* metabolism is further enhanced by a more detailed analysis of the reactions for which the flux balance analysis predicted zero flux. When clustering the measured gene expression profiles for the genes encoding the enzymes of these zero-flux reactions, a substantial number of genes showed consistent changes in gene expression along the time course, suggesting that the corresponding reactions are in fact active (Supporting information: Additional file 1). Striking examples include a large number of genes for vitamin B12 (cobalamin) biosynthesis, a group of ten genes involved in calcium-dependent antibiotic (CDA) biosynthesis, and three genes involved in ectoine biosynthesis (Supporting information: Additional files 2, 3, 4 and 5). Each of these cases provides important insights: the first one shows that vitamin B12 is likely to be produced by *S. coelicolor* under the growth conditions of our experiment, even if it is not essential due to the availability of cobalamine-independent enzymes (Martens et al., 2002). The second one highlights that CDA biosynthesis genes are coherently induced in expression during the metabolic switch, similar to undecylprodigiosin and actinorhodin and concordant with the results of the correlation analysis discussed above. This could indicate that this additional antibiotic compound is potentially also produced in phosphate starvation conditions, contrary to previous expectations (Kim et al., 2004). Finally, the case of ectoine biosynthesis genes suggests that this novel osmoprotectant metabolite is produced by *S. coelicolor*. This has in fact been experimentally confirmed recently (Kol et al., 2010). In each of these cases, the activity of the pathway was not predicted, based on the biological evidence incorporated in the stoichiometric model and the expected biomass composition, and the comparison of flux balance predictions and gene expression data indicated relevant modifications of our metabolic model. A complete list of genes that have zero predicted flux but show gene expression is included in the supporting information (Supporting information: Additional files 2, 3, 4, and 5).

Conversely, our model can be used to identify those genes that are predicted to be essential for growth (nonzero flux under all conditions), but show no or very low gene expression. There are 159 predicted essential genes in our model, which have a median log gene expression level of 7.47, compared to 6.83 for the non-essential genes and 4.66 for the negative controls. This indicates that on average the essential genes have a 60% higher expression than the non-essential genes. There is only one predicted essential gene with a detected median expression level below 5.0, compared to 23 non-essential genes with such low expression levels. This nonexpressed essential gene is panB (SCO2256), a 3-methyl-2-oxobutanoate hydroxymethyltransferase of pantothenate and coenzyme A biosynthesis, which has a maximum log expression signal of only 5.53. Its apparent non-expression can be due to insufficient hybridization of the gene–specific probes on the microarray, but it could also indicate the existence of another isoenzyme or additional metabolic pathways that would make this reaction redundant. In both of these cases, this gene might warrant further detailed study.

The observed good correlation between gene expression and predicted metabolic flux is not necessarily expected; expression levels can show little correlation to protein levels, enzyme activity and metabolic flux for many reasons (Akesson et al., 2004). It could be that the relationship between expression and flux is tighter in prokaryotes like *S. coelicolor*, than in multicellular eukaryotic model organisms (Wagner, 2005, 2007). However, we cannot exclude that the group of non-correlated genes contains not only reactions with constant flux, but also reactions with dynamic flux little correlation between gene expression and protein activity or metabolic flux. In a next step, it will be interesting to directly incorporate the gene expression information in the model, providing additional constraints on the maximum flux (Akesson et al., 2004; Colijn et al., 2009).

2.3 Conclusions

Our study demonstrates the ability of flux balance analysis to not only study classical steady-state conditions but also to predict microbial behavior in dynamic growth conditions provided that sufficiently detailed measurements of the changing growth conditions (nutrient uptake) and cellular objective (antibiotic production rate) are available. In combination with detailed gene expression information, these dynamic model predictions can help identifying potential new players in the metabolic switch, including putative new genes for antibiotic synthesis.

2.4 Methods

2.4.1 Transcriptomics

The gene expression dataset used in this study has been described in detail in (Nieselt et al., 2010). Briefly, *S. coelicolor* was cultivated in a phosphate limiting defined medium containing glucose as a carbon source and glutamate as a nitrogen as well as carbon source. Samples for transcriptomics and off-line analysis were taken every hour from 20 to 44 hours after inoculation (25 sample points), and subsequently every second hour from 46 to 60 hours after inoculation. Cell dry weight was measured on samples collected every third hour between 20 and 40 hours. The last sample, collected at the end of the fermentation (68 hours after inoculation), was used for analysis of remaining nutrients and total production levels of red and blue pigments. Only one sample was collected at each time point and no re-samplings were performed. Gene expression was measured on custom- made Affymetrix gene chips as described in (Nieselt et al., 2010). Expression data have been deposited in the GEO database under accession number GSE18489. Measurements for all known or predicted enzyme-coding genes were extracted and matched to the corresponding reactions in the constraint-based model.

2.4.2 Constraints-based genome-scale metabolic model reconstruction

A genome-scale stoichiometric metabolic model of *Streptomyces coelicolor* was reconstructed from different sources of data, including KEGG pathways, ScoCyc pathways, biochemistry textbooks, an extensive literature survey and available genome-scale models of other organisms. The initial stoichiometric matrix was generated based on KEGG and ScoCyc and manually curated to refine the *S. coelicolor*–specific parts of the metabolic network (e.g., antibiotic biosynthesis), to specify the correct reversibility constraints of reactions, and to add missing essential reactions. Missing essential reactions were identified iteratively; a minimum set of hypothetical reactions was added to the model if an essential metabolite could not be produced otherwise. Reversibility and essentiality of reactions were also compared to other published genome-scale models of *S. coelicolor* and other organisms (Borodina et al., 2005a; Reed et al., 2003; Jamshidi and Palsson, 2007; Oh et al., 2007). The resulting model is very similar to the model of Borodina et al. (Borodina et al., 2005a, 2008), and differs mainly in the more comprehensive inclusion of antibiotic pathways.

In the final curated model, one lumped reaction is added to produce the biomass of the cell. Information about biomass composition and growth and non-growth associated ATP maintenance were taken from Borodina et al. (Borodina et al., 2005a) and Ingraham et al. (Ingraham et al., 1983) and complemented with literature information (Shahab et al., 1996; Zuneda et al., 1984). Some of the biomass precursor biosynthesis reactions are also lumped reactions, e.g. protein translation, and were specified according to the literature and published genome-scale models (Borodina et al., 2005a). The full model in SBML format is available in the supporting information (Supporting information: Additional file 2). Analysis of the model was based on standard flux balance analysis (FBA) to predict optimal *in silico* growth and metabolic flux distribution using the COBRA tool (Becker et al., 2007). Uptake fluxes for metabolites not available in the medium were set to zero, while metabolic by-products were always allowed to leave the metabolic system. Observed nutrient uptake rates from the fermenter culture used for the transcriptome analysis were used to define the constraints of nutrients uptake for the model (input function). The objective function was defined as maximizing the growth rate. Beginning at 34 hours, we dynamically varied the biomass composition by adding increasing amounts of antibiotics, based on the observed antibiotics production rate.

2.4.3 Comparing transcriptome data and predicted flux

Our computational model contains 643 metabolites and 1015 reactions: 747 reactions for metabolite biosynthesis and degradation, 152 transport reactions, and 116 additional input and output constraints of the system. 666 reactions were annotated as enzyme–catalyzed reactions and could be matched to an enzyme–coding gene. Some reactions were annotated as potentially catalyzed by more than one gene and some genes catalyze more than one reaction. If one gene catalyzes multiple reactions, we matched its expression profile to the reaction with the maximum predicted flux, hypothesizing that this reaction will dominate the expression behavior. In total, 789 genes are assigned to 666 enzymatic reactions. Of these, 558 genes are predicted to have non-zero flux (the remaining 231 genes are not used for biomass production according to the model). Out of these 558 genes, 9 genes were involved in cell maintenance with constant flux and zero standard deviation; these were excluded from the further analysis. In total we therefore considered 549 enzyme–coding genes with non-zero predicted flux. For each of these genes, we compared the predicted flux profile and the observed gene expression levels using Pearson's correlation, testing whether gene expression was indeed upregulated when a much higher flux through a particular reaction was required at a certain growth phase.

2.5 Acknowledgements

We are very grateful to the STREAM consortium (funded by ERA-NET SySMO [GEN2006-27745-E/SYS]: (P-UK-01-11-3i) and the Research Council of Norway [project no. 181840/I30]) for providing the expression data prior to publication. The STREAM consortium is an international project funded by a SysMO grant of the ERASysBio framework http://www.erasysbio.net studying "Global metabolic switching in *Streptomyces coelicolor*". Members of the consortium include Florian Battke, Sven Even Borgos, Per Bruheim, Nigel Burroughs, Gregory Challis, Lubbert Dijkhuizen, Trond E Ellingsen, William Gaze, Alexander Herbig, Sunniva Hoel, yvind Jakobsen, Ritsert C. Jansen, Brent Kiernan, Preben Krabben, Roxane Legaie, Juan Francisco Martín, Jay Moore, Kay Nieselt, Walid Omara, David Rand, Jens Reuther, Antonio Rodríguez-García, Hvard Sletta, Maggie Smith, Louise Thomas, Alexander Wentzel, David Wild, Wolfgang Wohlleben, Arouna Woukeu, Anders Overby, and the authors https://www.wsbc.warwick.ac.uk/groups/sysmopublic/. MEM was funded by a 4 × 4 Ubbo Emmius scholarship and ET by a Rosalind Franklin Fellowship, both from the University of Groningen. RB is supported by an NWO-Vidi fellowship.

2.6 Authors' contributions

EMHW, ET and RB designed and coordinated the study. MTA carried out the modelling and drafted the manuscript. The STREAM consortium provided the expression data prior to publication. MTA and MEM integrated the model and expression data. DAH, ET and RB interpreted the results. EMHW, ET and RB revised the manuscript. All authors read and approved the final manuscript.

2.7 Supporting information

http://www.biomedcentral.com/1471-2164/11/202/additional/

Additional file 1: Expression clustering plots. PDF file depicting the expression clustering of 231 enzyme–coding genes for which the catalyzed reaction had zero predicted flux at all time points of the flux balance analysis of our model. The majority of genes are members of clusters that show highly consistent dynamics across the time course, e.g. the purple, red and pink clusters, indicating that they are indeed expressed and the corresponding reactions likely to be active.

Additional file 2: Stoichiometric metabolic model. SBML file describing the metabolic model of *Streptomyces coelicolor*.

Additional file 3: Table of metabolites. Excel table defining all metabolites used in the metabolic model.

Additional file 4: Table of reactions. Excel table defining all reactions used in the metabolic model.

Additional file 5: Table of zero-flux reactions. Excel table of reactions that show consistent zero predicted flux, including their membership in the expression clusters depicted in Additional files 2.

Published as: Medema MH*, **Alam MT***, Heijne W, Berg MVD, Müller U, Trefzer A, Bovenberg RAL, Breitling R, Takano E (2011): *Genome-wide gene expression changes in an industrial clavulanic acid overproduction strain of* Streptomyces clavuligerus **Microb. Biotechnol. 4(2):300–5.** *Equal contribution

Chapter 3

Genome-wide gene expression changes in an industrial clavulanic acid overproduction strain of *Streptomyces clavuligerus*

Abstract

To increase production of the important pharmaceutical compound clavulanic acid, a β–lactamase inhibitor, both random mutagenesis approaches and rational engineering of Streptomyces clavuligerus *strains have been extensively applied. Here, for the first time, we compared genome-wide gene expression of an industrial* S. clavuligerus *strain, obtained through iterative mutagenesis, with that of the wild-type strain. Intriguingly, we found that the majority of the changes contributed not to a complex rewiring of primary metabolism but consisted of a simple upregulation of various antibiotic biosynthesis gene clusters. A few additional transcriptional changes in primary metabolism at key points seem to divert metabolic fluxes to the biosynthetic precursors for clavulanic acid. In general, the observed changes largely coincide with genes that have been targeted by rational engineering in recent years, yet the presence of a number of previously unexplored genes clearly demonstrates that functional genomic analysis can provide new leads for strain improvement in biotechnology.*

3.1 Introduction

Streptomyces clavuligerus is an important industrial microorganism, which produces the beta lactam antibiotic cephamycin C (Martín and Liras, 1989) and the β–lactamase inhibitor clavulanic acid (Saudagar et al., 2008). Clavulanic acid is produced worldwide on a large scale, and co-formulated with amoxicillin in Augmentin (Brogden et al., 1981). Two biotechnological strain optimization approaches have been utilized to increase the production of clavulanic acid by the bacterium: rational metabolic engineering and iterative optimization through random mutagenesis and screening.

In the rational approach, a specific gene is knocked out or overexpressed to divert metabolic fluxes towards the antibiotic biosynthetic pathways. Arguably, the best example comes from the work of Li and Townsend (Li and Townsend, 2006), who re-engineered the *S. clavuligerus* glycolytic pathway by constructing a deletion mutant of the glyceraldehyde 3-phosphate dehydrogenase gene *gap1* to increase the pool of the clavulanic acid precursor glycerol 3-phosphate (G3P). This doubled clavulanic acid production compared to the wild-type. Overexpression of the regulatory proteins CcaR and ClaR also led to clavulanic acid overproduction (Hung et al., 2007), and the two strategies have recently been combined successfully in a single strain (Jnawali et al., 2010).

Yet, most or even all production strains that are used in industry have been obtained by classical strain improvement (Adrio and Demain, 2006), based on mutagenesis with mutagens such as nitrosoguanidine (NTG). Little is known about the exact genetic changes through which high production titers are achieved in these mutants.

Recently, we published the genome sequence of *S. clavuligerus* ATCC 27064 (Medema et al., 2010). We now employed this information to perform a genome-wide transcriptome study on an industrial production strain, which has been generated from the ATCC 27064 type strain (Higgens and Kastner, 1971) by several iterations of mutagenesis and screening, and produces clavulanic acid at levels approximately 100× that of the wild-type. The majority of observed changes consist of increased transcript levels of the antibiotic biosynthesis gene clusters. These observations are in agreement with flux-balance analysis (FBA) predictions using a constraint-based genome-scale metabolic model constructed for *S. clavuligerus*. However, we also detected some potentially crucial transcript level changes in primary metabolism that could contribute to the increased production of clavulanic acid by redirection of fluxes, mimicking strategies utilized in rational approaches.

3.2 Results and discussion

3.2.1 Increased transcription of secondary metabolite biosynthesis gene clusters in strain DS48802

Comparing gene transcript levels of *S. clavuligerus* wild-type and DS48802 strains during stationary phase using microarrays, almost all genes ranking high in a differential transcriptome analysis belong to the complete clavulanic acid/cephamycin C supercluster, which is significantly overexpressed (between twofold and eightfold) in the DS48802 strain compared to the wild-type. Interestingly, the pathway-specific regulator genes *claR* and *ccaR* are also overexpressed in DS48802. They are located within the same supercluster and their products have been shown to regulate it positively (Alexander and Jensen, 1998; Paradkar et al., 1998).

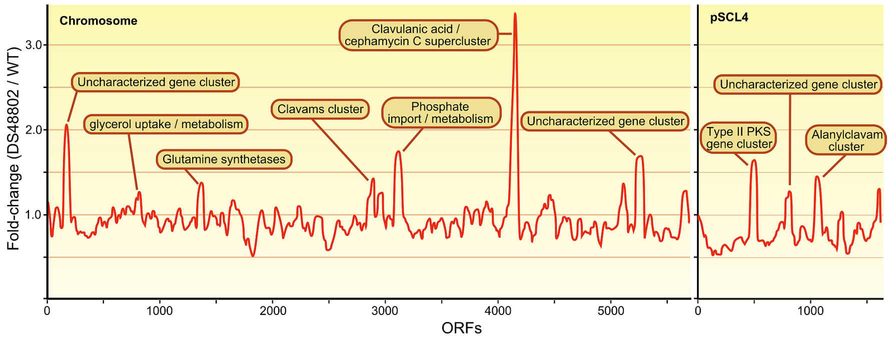

Figure 3.1: ***Differential gene expression in* S. clavuligerus *DS48802 and ATCC 27064.*** *Sliding window plot (size = 50) of the difference in gene expression between* S. clavuligerus *DS48802 and wild-type ATCC 27064. Key upregulated operons or genes at the peaks are noted in the figure. See supplementary Table S2 (section 3.5) for description of the ten gene clusters shown. For gene expression analysis, cultivations were performed in shake flasks directly inoculated with spore suspensions at 28°C and 280 r.p.m. The semi-synthetic growth medium used consisted of 30 (g/l) glycerol, 5 (g/l) wheat gluten, 3.5 (g/l) asparagine monohydrate, 1.5 (g/l) L-lysine, 0.7 (g/) KH2PO4, 0.3 (g/l) MgSO47H2O, 0.2 (g/l) CaCl22H2O, 0.2 (g/l) FeSO47H2O, 10 (g/l) MOPS, 0.1 (ml/l) Basilodon and 1 (ml/l) trace ele-*

ments solution at pH 7.0. The trace element solution consisted of 20.4 (g/l) H2SO4, 50 (g/l) citric acid H2O, 16.75 (g/l) ZnSO47H2O, 1.6 (g/l) CuSO4. 5 H2O, 1.5 (g/l) MnCl24H2O, 2 (g/l) H3BO3 and 2 (g/l) Na2MoO42H2O. After 70 h of cultivation, the cells were harvested by centrifugation, treated with RNAprotect (Qiagen) and directly frozen with liquid nitrogen and stored at –80°C. To isolate total RNA, the frozen mycelium was ground in a mortar, resuspended in TE buffer with 5 (mg/l) lysozyme and incubated for 5 min at room temperature. RNA isolation and purification were performed using phenol extraction (TRIzol reagent, Invitrogen) and RNeasy Kit (Qiagen). The RNA was quantified by measuring the absorbance at 260 nm. Biotinylated cDNA was prepared after fragmentation according to the standard Affymetrix protocol using GC rich (average 72%) primers from 10 μg total RNA. For hybridization, 5 μg and 7 μg biotinylated cDNA were used per Affymetrix gene Chip. Microarray data have been deposited at Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/) under accession number GSE24033. Flux-balance analysis was performed using a recently published genome-scale metabolic model of S. clavuligerus *(Medema et al., 2010). In this study, we slightly changed our objective function and included both clavulanic acid and cephamycin C biosynthesis pathways. We dynamically changed the antibiotic concentration in the biomass composition based on experimental observations of Romero and colleagues (Romero et al., 1984) and optimized the objective function for different concentrations of each antibiotic. Among the 785 genes that the model contains, 497 genes showed non-zero flux for at least one antibiotic concentration. We calculated Spearman correlation of fluxes of each reaction with increasing antibiotic concentrations. If an enzyme was involved in multiple reactions, we assigned the flux which had the highest r2.*

Additionally, the clavams gene cluster (*cvm1245* and *cas1*) and the 'paralogous' alanylclavam cluster (*orfABCD* and *ceaS1/pah1/bls1/oat1*) are significantly overexpressed (Figure 3.1), as is the two-component system involving Cvm7p (SCLAV_p1079 – p1080) that induces expression of the alanylclavam cluster (Tahlan et al., 2007). This suggests the presence of a regulatory mechanism common to all these clusters. *CcaR* is an unlikely candidate for such a common regulatory factor, as it does not appear to control the paralogous cluster (Tahlan et al., 2004); the pleiotropic regulator AdpA (SCLAV_1957) is a more likely candidate, as it is known to induce clavulanic acid expression (López-García et al., 2010) and its gene is transcribed almost 2.5 times stronger in DS48802.

In contrast to, for example, the intrachromosomal amplification of the kanamycin biosynthesis gene cluster in *Streptomyces kanamyceticus* (Yanai et al., 2006), hybridization of *S. clavuligerus* DS48802 genomic DNA to the microarrays revealed no amplifications of genes or gene clusters (data not shown). The overproduction that we observe therefore appears to be caused by transcriptional (and post-transcriptional) changes only.

3.2.2 Flux-balance analysis of increased clavulanic acid production correlates well with transcriptomic data

While many changes have occurred during the generation of the DS48802 strain through mutagenesis, it is important to note that changes in transcription levels of many genes may be due to random mutations that have no impact on antibiotic biosynthesis. In order to assess which changes could be causatively linked to antibiotic overproduction, we computationally predicted the metabolic fluxes during antibiotic overproduction, using a constraints-based genome-scale metabolic network model of *S. clavuligerus* (Medema et al., 2010). We dynamically modeled the metabolic flux changes during increased production of clavulanic acid and cephamycin C with different rates of antibiotic production relative to the biomass, based on the experimental observations of Romero et al. (Romero et al., 1984). Interestingly, the computational predictions made through dynamic FBA are largely in line with the observed expression changes. Eighty-seven genes were predicted to be upregulated (positively correlated with antibiotic production; $r > 0.6$) and 129 genes were predicted to be downregulated (negatively correlated; $r < -0.6$) at increasing antibiotic production levels. Fourty per cent (15/37) of the genes that actually showed increased transcript levels (fold change > 2) were also predicted to do so according to FBA, and these include all genes encoding key biosynthetic enzymes known to be involved in clavulanic acid and cephamycin C biosynthesis. Even though 40% does not appear to be a very large percentage, these predictions are statistically very significant according to a Fisher's exact test (P = 0.0005; see Supporting information: Table S1, section 3.5). One should note that FBA predicts the flux for every reaction, not for every gene product, as multiple gene products can be involved in a single reaction. Therefore, in these cases the same flux was assigned to all gene products involved in that particular reaction, which is not necessarily the case in the actual gene expression, as only a single homologue or isoenzyme could be actively performing the reaction and thus would be differentially expressed. Out of the 47 different enzymatic reactions predicted to be upregulated (associated with 87 genes), 26 (55%) have at least one gene linked to them, which showed increased transcript levels. As both our FBA and gene transcript analysis pointed to an increased expression of the clavulanic acid and cephamycin core biosynthesis genes, we suggest that this is a crucial change required for antibiotic overproduction in this strain. Moreover, as the FBA indicated that the absolute fluxes required for high production of the secondary metabolites are minor compared to other fluxes involved in maintenance and cellular growth, a complete redirection of primary metabolism appears not to be necessary for overproduction.

3.2.3 Gene expression changes in primary metabolism

Nonetheless, because clavulanic acid is synthesized from the precursors G3P and L-arginine, which play important roles in primary metabolism, specific changes in the primary metabolism of DS48802 could have occurred during the various random mutagenesis rounds, so that the intracellular pools of these intermediates are increased.

Indeed, glycerol uptake and metabolism (SCLAV_0631 – 0632) and (SCLAV-_0877 – 0879) is clearly upregulated over twofold in DS48802, indicating an improved utilization of glycerol as a carbon source as well as increased production of the clavulanic acid precursor G3P (Figure 3.2). Moreover, the aconitase and citrate synthase from the citric acid cycle appear to be downregulated. A likely explanation for this is that the carbon flux from G3P in this direction is reduced and is partly redirected to clavulanic acid biosynthesis. This situation is remarkably similar to the result of the rationally constructed *gap1* deletion that blocked G3P conversion into 1,3-bisphosphoglycerate, thus improving clavulanic acid biosynthesis by increasing the intracellular G3P pool (Li and Townsend, 2006). However, an advantage of the situation in DS48802, which seems to have an incomplete downregulation of the flux, could be that a considerable pool of acetyl-CoA is maintained, e.g. for the biosynthesis of ornithine from glutamate. DS48802 also seems to avoid the potential negative effects of a complete deletion of the aconitase and citrate synthase genes: a complete absence of these enzyme activities could lead to acidogenesis with negative consequences for secondary metabolite production as shown by Viollier and colleagues (Viollier et al., 2001a,b). Moreover, DS48802 still seems to be able to synthesize alpha-ketoglutarate (a co-substrate required for clavaminic acid biosynthesis; (Salowe et al., 1990)), while achieving the benefits of higher acetyl-CoA and/or G3P pools that have made these genes attractive targets for rational engineering to improve antibiotic production (Viollier et al., 2001a). A potentially important observation that we cannot explain yet from the current data are the differential transcript level changes of the two pyruvate kinase isoenzyme genes, one being downregulated (SCLAV_4329) and the other being upregulated (SCLAV_1203).

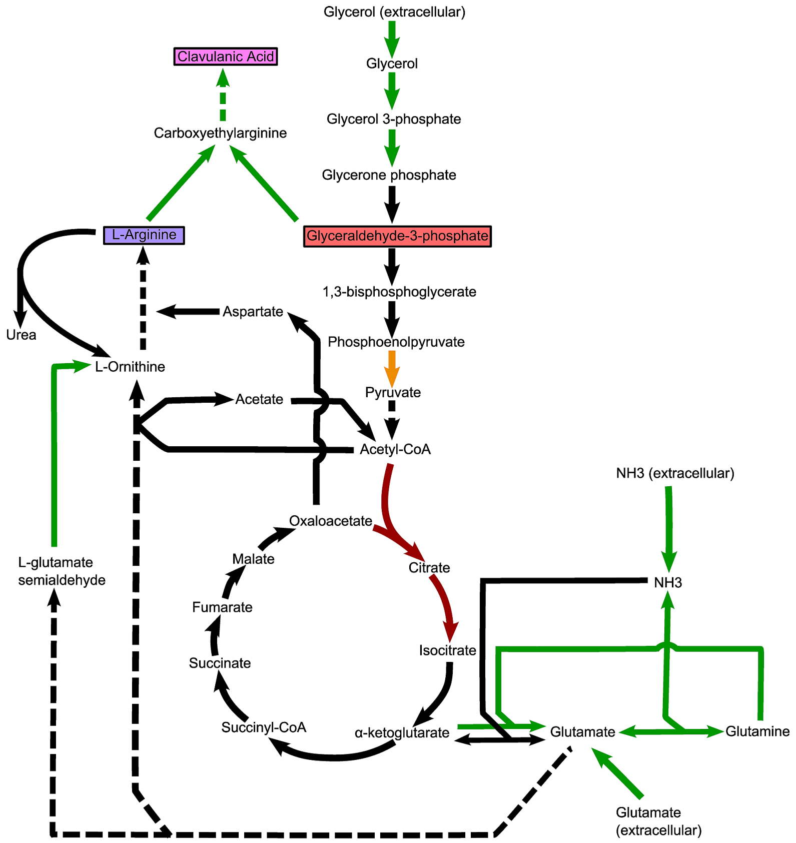

Figure 3.2: ***Changes in* S. clavuligerus *primary and secondary metabolism affecting clavulanic acid production.*** *Changes in gene expression in* S. clavuligerus *DS48802 compared to the wild-type ATCC 27064 projected onto a metabolic map. Green arrows represent reactions catalyzed by genes expressed over twofold higher in DS48802 than in the wild-type. Red arrows represent reactions catalyzed by genes expressed over twofold lower in DS48802. The orange arrow represents the reaction catalyzed by pyruvate kinase, for which two isoenzymes exist which have changed in expression differently, one being downregulated (SCLAV_4329) and the other being upregulated (SCLAV_1203). Black arrows represent unchanged steps; solid arrows represent single biosynthetic steps; and dashed arrows represent multiple steps.*

We also observed a remarkable upregulation of glutamine synthetases I and II (SCLAV_1416 and SCLAV_1431), glutamate synthetases (SCLAV_1231) and glutamate importers (SCLAV_4660 – 4663). Glutamate can serve as a source for biosynthesis of the clavulanic acid precursor arginine. This conversion takes place through the urea cycle involving ornithine as an intermediate (Rodríguez-García et al., 2000), addition of which to the medium has been shown to strongly enhance clavulanic acid biosynthesis (Chen et al., 2003). Probably to overcome nitrogen and phosphate limitations, genes encoding the transporters for ammonia (SCLAV_4534) and phosphate (SCLAV_3166 – 3169) are also observed to be more highly transcribed in DS48802. This may be caused by the increased expression of the pathway-specific activator genes phoU (SCLAV_3220, (Ghorbel et al., 2006)) and glnB (SCLAV_4535, (Drepper et al., 2003)), which both show over twofold increased transcription.

3.3 Conclusions

Our data show that a strain improvement program by random mutagenesis and screening has caused gene transcript changes in both primary and secondary metabolism. The overlap with results obtained by rational metabolic engineering through *claR/ccaR* overexpression and *gap1* deletion is intriguing. New leads from transcript changes observed in this study, such as the increased transcription of glutamine and glutamate synthetase genes, and of those encoding ammonium and phosphate transporters, can now be combined to rationally design novel high-producer strains. This approach might avoid the introduction of unwanted adverse effects from random mutagenesis, and provide strains suited for industrial application in a more efficient way. In this manner, functional genomics allows two key strategies applied in biotechnology – random mutagenesis and rational engineering – to become increasingly complementary.

3.4 Acknowledgements

This work was supported by the Dutch Technology Foundation STW, which is the applied science division of NWO, and the Technology Programme of the Ministry of Economic Affairs [STW 10463]. RB is supported by an NWO-Vidi fellowship, and ET by a Rosalind Franklin Fellowship, University of Groningen. We thank Bert Koekman for providing the classically improved strain DS48802. We thank Hilde Huininga and Hildegard Menke for technical assistance with DNA and RNA isolations.

3.5 Supporting information

http://onlinelibrary.wiley.com/doi/10.1111/j.1751-7915.2010.00226.x/full

Table S1: Statistical significance and comparison of transcriptional changes with FBA predictions using different cut-offs.

Table S2: Description and gene expression data of gene clusters upregulated in *S. clavuligerus* DS48802.

Published as: **Alam MT**, Merlo ME, Takano E, Breitling R (2010): *Genome-based phylogenetic analysis of* Streptomyces *and its relatives.* **Molecular Phylogenetics and Evolution 54 (2010) 763–772**

Chapter 4

Genome-based phylogenetic analysis of *Streptomyces* and its relatives

ABSTRACT

Motivation: Streptomyces *is one of the best-studied genera of the order* Actinomycetales *due to its great importance in medical science, ecology and the biotechnology industry. A comprehensive, detailed and robust phylogeny of* Streptomyces *and its relatives is needed for understanding how this group emerged and maintained such a vast diversity throughout evolution and how soil-living mycelial forms (e.g.,* Streptomyces *s. str.) are related to parasitic, unicellular pathogens (e.g.,* Mycobacterium tuberculosis*) or marine species (e.g.,* Salinispora tropica*). The most important application area of such a phylogenetic analysis will be in the comparative re-annotation of genome sequences and the reconstruction of* Streptomyces *metabolic networks for biotechnology.*

Methods: *Classical 16S-rRNA-based phylogenetic reconstruction does not guarantee to produce well-resolved robust trees that reflect the overall relationship between bacterial species with widespread horizontal gene transfer. In our study we therefore combine three whole genome-based phylogenies with eight different, highly informative single-gene phylogenies to determine a new robust consensus tree of 45* Actinomycetales *species with completely sequenced genomes.*

Results: *None of the individual methods achieved a resolved phylogeny of* Streptomyces *and its relatives. Single-gene approaches failed to yield a detailed phylogeny; even though the single trees are in good agreement among each other, they show very low resolution of inner branches. The three whole genome-based methods improve resolution considerably. Only by combining the phylogenies from single gene-based and genome-based approaches we finally obtained a consensus tree with well-resolved branches for the entire set of* Actinomycetales *species. This phylogenetic information is stable and in-*

formative enough for application to the system-wide comparative modeling of bacterial physiology.

Keywords: Phylogeny; *Streptomyces*; *Actinomycetales*; Rank order; rRNA

4.1 Introduction

Streptomyces species are among the best-studied and best-characterized bacteria due to their significant role for medical science, ecology and the biotech industry (Lechevalier and Lechevalier, 1967; Embley and Stackebrandt, 1994; Bentley et al., 2002). *Streptomyces* have a particularly complex secondary metabolism, which produces a large collection of biologically useful compounds. Most importantly, they are employed at large industrial scale in the production of most of the available antibiotics applied in human and veterinary medicine, as well as a large number of anti-parasitic agents, herbicides, immuno-suppressants and several enzymes important in the food and other industries (Bentley et al., 2002; Cerdeño-Tárraga et al., 2003; Hopwood, 2007).

The genus *Streptomyces* is taxonomically located in the diverse bacterial order *Actinomycetales*. This group is characterized by an astonishing diversity in terms of morphology, ecology, pathogenicity, genome size, genomic G+C content, and the number of coding sequences in the genome (Embley and Stackebrandt, 1994; Ventura et al., 2007; Hopwood, 2007). Morphologically, some species are rod-shaped or coccoid, while others form fragmenting hyphae or a branched mycelium (Ventura et al., 2007). Spore formation is also very common in *Actinomycetales* but it is not ubiquitous (Ventura et al., 2007). Ecologically, some actinomycetes are soil–living bacteria, some are marine, are colonizing thermal springs (Barabote et al., 2009) or growing on gamma-irradiated surfaces (Phillips et al., 2002) or as plant root symbionts (Normand et al., 2007) and some are important pathogens of humans, animals and plants (Goodfellow and Williams, 1983; Castillo et al., 2002; Tokala et al., 2002). For instance, *Mycobacterium tuberculosis* causes tuberculosis (Cole et al., 1998), *Corynebacterium diptheria* infection results in diptheria (Cerdeño-Tárraga et al., 2003), *Propionibacterium acnes* is the agent of acnes (Leyden, 2001) and *Streptomyces scabies* causes potato scab (Takeuchi et al., 1996). *Streptomyces* species are found mostly in the soil where they live as saprophytes (Hopwood, 2007), but recently some species have been described in the rhizosphere of plant roots and in other plant tissue (Castillo et al., 2002; Tokala et al., 2002), isolated from leaf cutting ants (Kost et al., 2007) and also associated with marine sponge species (Zhang et al., 2008).

Considering the importance of *Streptomyces* and its relatives in terms of both biological behavior and metabolic products, it becomes essential to understand its evolutionary relationships to other species in the diverse *Actinomycetales* order. On the one hand, an evolutionary study may help to explain how *Streptomyces* emerged and adapted to the soil environment. On the other hand, information obtained from a well-resolved phylogeny can be used for the comparison of genome sequences, comparative genome re-annotation, and genome visualization. A robust phylogeny is central for ongoing efforts in many groups to reconstruct system-wide metabolic models of *Streptomyces* and related species (Borodina et al., 2005a), which are used for systematic strain-engineering in biotechnology.

Currently available phylogenies of the group are based on 16S rRNA, individual genes, or comparative genomic approaches (Embley and Stackebrandt, 1994; Takeuchi et al., 1996; Stackebrandt et al., 1997; Anderson and Wellington, 2001; Egan et al., 2001; Gao and Gupta, 2005; Chater and Chandra, 2006; Manteca et al., 2006). Such reconstructions tend to be relatively unstable and are not guaranteed to reflect the overall evolutionary history in a complex group with widespread horizontal gene transfer. To address this issue and to determine potential problematic areas in the single-gene phylogenies, we made use of whole-genome information. We not only reconstructed phylogenetic trees based on eight highly conserved single genes, including three rRNA sequences (*5S*, *16S*, *23S* rRNA) and five ubiquitous protein sequences (*isoleucyl tRNA synthetase*, *ribosomal protein S1*, *SecY*, *GTPase*, *DNA topoisomerase*), but also integrated three different phylogenetic reconstructions based on complete genome sequences, using gene content, gene order and gene concatenation analysis.

Our analysis is based on 45 species from eight different suborders of

Actinomycetales, including four genome sequenced *Streptomyces species* (Table 4.1). *Escherichia coli* and *Bacillus subtilis* are used as outgroups, and two distantly related *Actinobacteria, Bifidobacterium longum* NCC2705 and *Bifidobacterium adolescentis* ATCC 15703, serve as sub-outgroups.

Our results showed that single-gene approaches did not yield a resolved phylogeny of *Actinomycetales*. Resolution is much improved in the whole genome-based reconstructions: the phylogenetic trees are in good agreement with each other and better resolution of inner branches is achieved. We combine the information from the single-gene phylogenies and the whole genome-based approaches to reconstruct a final consensus tree. The resulting tree shows a detailed, well-supported phylogeny of the 45 *Actinomycetales* species with complete resolution of inner branches at species level, which was not achieved independently by any of the individual methods which generally show weakly resolved clusters. The consensus tree is in good agreement with the traditional taxonomic classification at the family and suborder level. The only major exception is *Kineococcus radiotolerans* SRS30216, which is placed in *Micrococcineae* rather than in *Frankineae* as reported previously (Lilburn and Garrity, 2004; Lee, 2006; Normand et al., 2007), in agreement with the most recent revision of the *Actinobacteria* (Zhi et al., 2009).

Species in the same color belong to the same family or suborders in traditional taxonomies. *Bifidobacteriales* are close relatives of *Actinomycetales*, and also belong to the order *Actinobacteridae*. They are used as sub-outgroups, together with the more distantly related outgroup species *Escherichia coli* and *Bacillus subtilis*.

4.2 Methods

In this study we combine single-gene and whole-genome approaches to obtain a completely resolved picture of the phylogeny of *Streptomyces* and their relatives.

Table 4.1: ***List of species considered in the study***

Suborder	Family	Species	Species kegg ID
Corynebacterineae	Corynebacteriaceae	Corynebacterium jeikeium K411	cjk
		Corynebacterium glutamicum ATCC 13032	cgl
		Corynebacterium diphtheriae NCTC 13129	cdi
		Corynebacterium efficiens YS-314	cef
		Corynebacterium glutamicum ATCC 13032_new	cgb
	Mycobacteriaceae	Mycobacterium leprae TN	mle
		Mycobacterium tuberculosis H37Ra	mra
		Mycobacterium bovis BCG str. Pasteur 1173P2	mbb
		Mycobacterium tuberculosis F11	mtf
		Mycobacterium ulcerans Agy99	mul
		Mycobacterium gilvum PYR-GCK	mgi
		Mycobacterium vanbaalenii PYR-1	mva
		Mycobacterium sp. KMS	mkm
		Mycobacterium smegmatis str. MC2 155	msm
		Mycobacterium avium 104	mav
		Mycobacterium sp. MCS	mmc
		Mycobacterium avium subsp. paratuberculosis	map
		Mycobacterium bovis AF2122/97	mbo
		Mycobacterium tuberculosis CDC1551	mtc
		Mycobacterium tuberculosis H37Rv	mtu
		Mycobacterium sp. JLS	mjl
	Nocardiaceae	Rhodococcus sp. RHA1	rha
		Nocardia farcinica IFM 10152	nfa
Frankineae	Acidothermaceae	Acidothermus cellulolyticus 11B	ace
	Frankiaceae	Frankia alni ACN14a	fal
		Frankia sp. EAN1pec	fre
		Frankia sp. CcI3	fra
Kineosporiineae	Kineosporiaceae	Kineococcus radiotolerans SRS30216	kra
Micrococcineae	Cellulomonadaceae	Tropheryma whipplei TW08/27	tws
		Tropheryma whipplei str. Twist	twh
	Microbacteriaceae	Clavibacter michiganensis subsp. michiganensis NCPPB 382	cmi
		Leifsonia xyli subsp. xyli str. CTCB07	lxx
	Micrococcaceae	Arthrobacter sp. FB24	art
		Arthrobacter aurescens TC1	aau
		Renibacterium salmoninarum ATCC 3209	rsa
Micromonosporineae		Salinispora arenicola CNS-205	sar
		Salinispora tropica CNB-440	stp
Propionibacterineae		Propionibacterium acnes KPA171202	pac
		Nocardioides sp. JS614	nca
Pseudonocardineae		Saccharopolyspora erythraea NRRL 2338	sen
Streptomycineae		Streptomyces avermitilis MA-4680	sav
		Streptomyces coelicolor A3(2)	sco
		Streptomyces scabies	ssc
		Streptomyces griseus	sgr
Streptosporangineae		Thermobifida fusca YX	tfu
Bifidobacteriales		Bifidobacterium longum NCC2705	blo
		Bifidobacterium adolescentis ATCC 15703	bad
Outgroup		Escherichia coli	eco
		Bacillus subtilis	bsu

4.2.1 Gene selection and single gene-based phylogeny

The majority of large scale phylogenetic reconstructions are currently based on rRNA sequences. Since rRNAs are essential, they are highly conserved throughout different species, lateral transfer is very rare, and their molecular size ensures they carry abundant evolutionarily informative sites. This makes rRNA extremely informative for phylogenetic analysis (Woese, 1987). For this study we included 3 different rRNA genes (*5S*, *16S* and *23S*) for single-gene analysis. For comparison, we included 5 large and broadly conserved protein sequences: *isoleucyl tRNA synthetase*, *ribosomal protein S1*, *DNA topoisomerase*, *SecY* and *GTPase*, involved in a variety of conserved cellular functions. The selection of these protein sequences was based on their level of conservation (Figure 4.1 A) and sequence length. Since each sequence position contains information on a very narrow range of evolutionary time, a larger number of independently evolving positions leads to better phylogenetic resolution (Olsen and Woese, 1993). Reconstructions based on larger molecules are also less affected by local non-random re-arrangements (Snel et al., 2005; Kunisawa, 2007).

Sequences were aligned separately by ClustalW (Thompson et al., 1994), sites containing gaps were removed, and 100 bootstrap replicates of each individual sequences alignment were generated. We used neighbor-joining (Saitou and Nei, 1987), Fitch-Margoliash (Fitch and Margoliash, 1967), maximum parsimony (Sourdis and Nei, 1988) and maximum likelihood (Felsenstein, 1981), all implemented in the PHYLIP package (Felsenstein, 2007) as alternative approaches for phylogenetic inference. Final trees for all individual sequences were built by using the CONSENSUS program using the majority-rule consensus approach on all bootstrap results (Felsenstein, 2007). The resulting trees show a conservative picture of the phylogeny, including only the best-supported relationships as resolved.

4.2.2 Whole genome-based phylogeny

For the genome-based approach we collected all coding sequences from the genome annotations of 45 *Actinomycetales* species and 4 outgroup species (Table 4.1) available in the National Center for Biotechnology Information database (http://www.ncbi.nlm.nih.gov/). Homologs were assigned using a reciprocal best hit strategy. We found 155 broadly conserved proteins, which were present in all species and were used for gene concatenation and gene order-based phylogenetic reconstruction. A third analysis was based on the entire gene content of the same 49 genomes.

Gene concatenation phylogeny

Phylogeny based on concatenated gene sequence data (Brown et al., 2001; Herniou et al., 2001; Ciccarelli et al., 2006) is an intuitive extension of the single-gene approach. Instead of focusing on one or a few genes, it uses the concatenated sequence of all conserved proteins (Figure 4.1 B). Hence, the approach is potentially more robust because of the greatly increased number of phylogenetic informative sites. The 155 ubiquitous protein sequences were individually aligned using ClustalW (Thompson et al., 1994) and, after removing all sites containing gaps, the 155 alignments were concatenated into one meta-alignment. The meta-alignment contained a total of 40333 phylogenetically informative sites. A complete list of all proteins included is provided in the Supporting information. For tree building we again used neighbor-joining, Fitch-Margoliash and maximum parsimony, with 100 bootstrap replicates. Due to computational constraints, maximum likelihood was not used for this part of the study.

Gene order phylogeny

Quite independent of the gene sequences, the physical order of genes along the genome is phylogenetically highly informative, and earlier studies have already confirmed that phylogenetically related genomes clearly have similar gene order (Koonin et al., 2000; Rokas and Holland, 2000; Snel et al., 2005; Kunisawa, 2007). An important advantage is that this approach is largely independent of sequence alignments and will not be affected by misalignments that often lead to wrong tree topologies. We again considered all 155 ubiquitous genes and then compared their left and right neighbors for all pairs of genomes (Figure 4.1 C). We calculate a similarity score for each species pair, determining how many neighbors are shared (maximum score = 2 * 155). To build a tree from the resulting similarity matrix we used three different distance methods: neighbor-joining (Saitou and Nei, 1987), Fitch-Margoliash (Fitch and Margoliash, 1967) and Kitch (Felsenstein, 2007) with 100 bootstrap replicates.

Gene content phylogeny

The two genome-based approaches described so far consider only those genes that are universally conserved. Those molecules which are conserved only in sub-clades are ignored. To overcome this limitation, we used gene content analysis (Figure 4.1 D), which is a comprehensive way of phylogenetic inference which takes into account conservation distributed all over the genome (Gibbon and House, 1999; Snel et al., 1999; Lin and Gerstein, 2000; Montague and Hutchison, 2000; Daubin et al., 2002; Huson and Steel, 2004; Henz et al., 2005). Here we use a very simple and robust way to implement this phylogenetic inference strategy: For each pair of species we calculated the percentage of genes that have a reciprocal best hit (homolog), relative to the number of genes in the longer of the two genomes.

The resulting similarity matrix was used for phylogenetic reconstruction using neighbor-joining, Fitch-Margoliash, and Kitch with 100 bootstrap replicates.

4.2.3 Rank order

A phylogenetic tree visualizes the evolutionary relationships among organisms by grouping them in different clades. However, for some applications, it is more convenient to also obtain a linear ordering of species according to their similarity to a target species. Such a ranking of organisms is particularly useful for genome visualization, but also for determining the proper weighting of species in comparative metabolic network reconstruction. We implemented three different ways of ranking the species, based on average rank, median rank and rank product (Breitling et al., 2004). The ranking of species based on their phylogenetic distances to *Streptomyces coelicolor*, our main target species, was obtained from all eight single gene-based phylogenies and from all three genome-based methods.

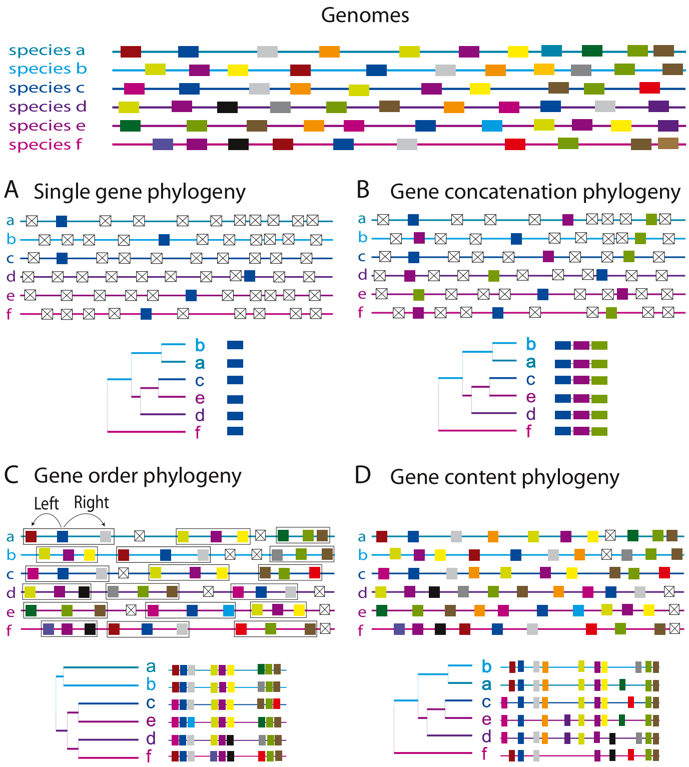

Figure 4.1: *Cartoon of the independent sources of information used to resolve the phylogeny of 45* Actinomycetales *species. Six different genomes are illustrated in the cartoon. Each box represents a gene and boxes with the same color represent homologous genes in different genomes. Four different independent sources of information have been used for the phylogenetic reconstruction: (a) Single-gene phylogeny considers only individual, highly conserved genes in different genomes. The blue boxes represent one selected universally conserved gene, while crossed boxes indicate those genes which are not considered in the tree reconstruction. (b) Gene concatenation phylogeny uses the concatenated sequence of all the universally conserved proteins in the genomes; crossed boxes indicate genes that are not taken into consideration because they are not ubiquitous. (c) Gene order phylogeny takes into account the physical order of genes: by comparing the identity (not the sequence) of the left and the right neighbors of each ubiquitous gene, a similarity matrix is obtained and used to build the phylogenetic tree; crossed boxes indicate genes that are neither ubiquitous nor neighbors of ubiquitous genes. (d) Gene content phylogeny takes into account conservation through all genomes: by calculating the percentage of homologs between pairs of species relative to the total number of genes, we obtain a similarity matrix which is used for phylogenetic reconstruction. Crossed boxes indicate that genes that occur only in single species.*

4.3 Results

Single gene-based phylogeny

As expected, all single-gene trees are rather poorly resolved. The tree based on 5S rRNA gene gives the most unresolved evolutionary picture, due to the short sequence: most of the organisms of *Actinomycetales* belonging to the same family are grouped in one cluster, but at the suborder level no resolution is shown (Figure 4.2 A). The phylogenetic trees based on 16S rRNA (Figure 4.2 B) and 23S rRNA (Figure 4.2 C) sequence genes are more resolved at the suborder level [Supporting information: Phylogenetic trees with bootstrap values], as is the case for the highly conserved protein trees [Supporting information: Protein sequence trees]. While the overall evolutionary structure is the same for all 8 consensus trees, there are a number of poorly resolved relationships, where various single-gene trees disagree.

For instance, in the 23S rRNA tree two families of the suborder *Frankineae* (*Acidothermaceae* and *Frankiaceae*) are clustered with the suborder *Streptomycineae* and a member of *Kineosporiaceae*, *Kineococcus radiotolerans* SRS30216, is merged with the suborder *Micrococcineae*. In the protein trees, *Kineococcus radiotolerans* SRS30216 is placed in the suborder *Micrococcineae* or unresolved. In the 16S rRNA tree the suborders *Frankineae* and *Streptomycineae* are shown in two unresolved independent branches. In contrast, in some of the protein trees, *Frankineae* is clustered with *Streptomycineae*.

Thermobifida fusca YX of the suborder *Streptosporangineae* is clustered with two families of *Frankineae* in the 16S rRNA tree, but with *Saccharopolyspora erythraea* NRRL 2338 of the suborder *Pseudonocardineae* in the 23S rRNA tree.

The placement of the suborder *Micromonosporineae* is not clearly resolved in the rRNA-based trees: in the 5S rRNA and 23S rRNA tree their placement is unresolved, while in 16S rRNA *Micromonosporinea* are with suborder *Frankineae*.

The families *Corynebacteriaceae*, *Mycobacteriacae* and *Nocardiaceae* of suborder *Corynebacterineae* which should be grouped in a monophyletic cluster, as indicated by Casanova and Abel (Casanova and Abel, 2002) fall in three separate unresolved branches instead in the rRNA trees, while the five protein

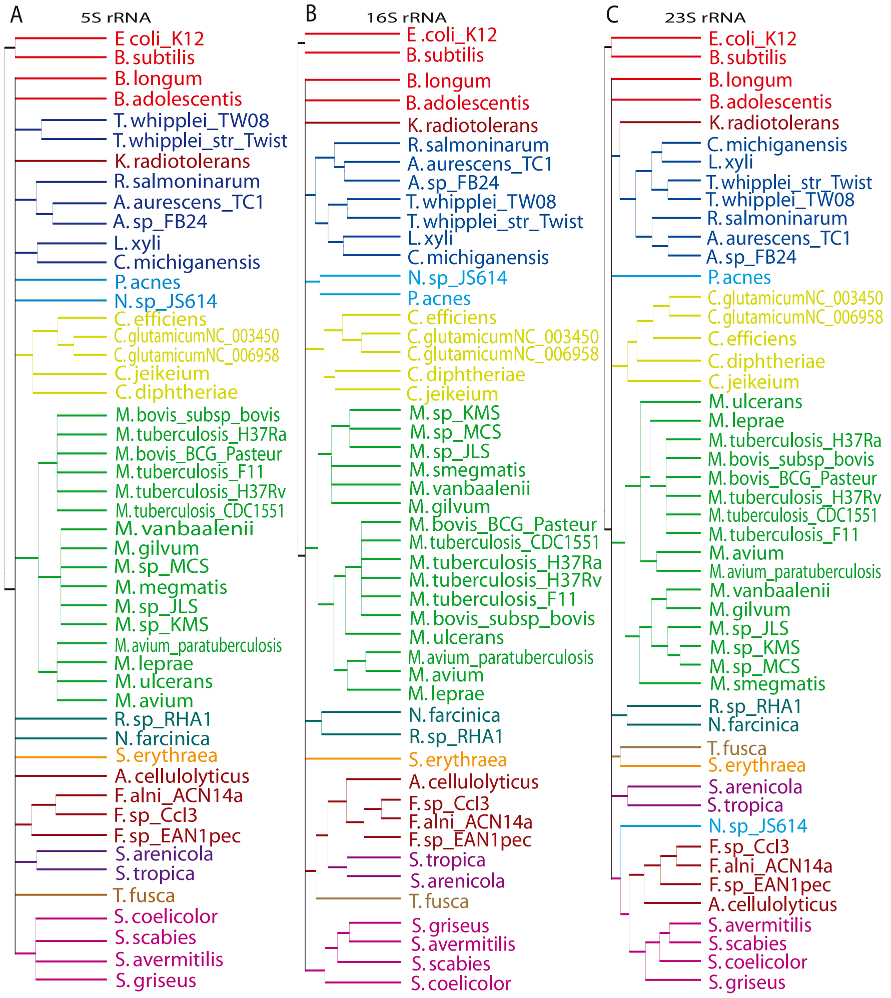

Figure 4.2: *Consensus trees based on rRNA sequences, (A) 5S rRNA, (B) 16S rRNA, (C) 23S rRNA. Initial trees were reconstructed using four tree inference approaches (NJ, Fitch, ML, PARS) with 100 bootstrap replicates each and combined by majority-rule consensus to include only well-supported branches. Organisms having the same colors are members of the same suborder in the taxonomical classification of NCBI.*

sequence trees show *Corynebacteriaceae* plus *Mycobacteriacae* plus *Nocardiaceae* as a monophyletic group as expected (Casanova and Abel, 2002), while other subgroups mostly remain unresolved.

Most species of *Mycobacteriacae* are well resolved, but the phylogeny among the six pathogenic species within the *Mycobacterium suborder* [*Mycobacterium tuberculosis* F11, *Mycobacterium tuberculosis* H37Rv, *Mycobacterium tuberculosis* H37Ra, *Mycobacterium tuberculosis* CDC1551, *Mycobacterium bovis* BCG str. Pasteur 1173P2, *Mycobacterium bovis* AF2122/97] is very poorly resolved: the six species form a single, completely unresolved branch in all rRNA consensus trees and in four protein-based trees.

When the tree is rooted on *E. coli* and *B. subtilis*, the sub-outgroups *Bifidobacterium longum* NCC2705 and *Bifidobacterium adolescentis* ATCC 15703 are placed on an unresolved branch among the actinomycetes.

Thus, from the analysis of three rRNA genes and eight proteins we could not obtain a completely resolved, well-supported picture of *Actinomycetales* phylogeny. To overcome this problem we moved onto whole genome-based phylogenetic reconstruction methods.

Gene concatenation phylogeny

The phylogenetic tree based on the concatenation of 155 protein sequences highly conserved among actinomycetes gives better resolution than single molecule based phylogeny (Figure 4.3 A). It shows excellent congruence with the 16S and 23S rRNA trees (Figure 4.2 A-C) for all major taxonomic groupings, confirming the validity of the method. One of the more resolved branches contains the families *Corynebacteriaceae*, *Mycobacteriacae* and *Nocardiaceae*, which are now clustered in a monophyletic group as expected (Koonin et al., 2000). The cluster containing the six pathogenic species of *Mycobacterium*, which was completely unresolved in the single gene-based phylogeny, is instead well resolved here in agreement with the taxonomic classifications in the NCBI database. Three families of the suborder *Micrococcineae* [*Cellulomonadaceae*, *Microbacteriaceae*, *Micrococcaceae*] are grouped together in one branch and individual species are correctly related to their traditional taxonomic family members. When the tree is rooted on *E. coli* and *B. subtilis*, *Bifidobacterium longum* NCC2705 and *Bifidobacterium adolescentis* ATCC 15703 are clearly present as well-resolved sub-outgroups.

An interesting agreement with the single-gene phylogenies is the grouping of the two families *Frankineae* and *Streptomycineae* with a member of *Streptosporangineae*, *Thermobifida fusca* YX. Surprisingly, *Kineococcus radiotolerans* SRS30216 still does not branch with any *Frankineae* species, in contrast to traditional phylogenies (Lee, 2006; Normand et al., 2007; Garrity et al., 2004).

Gene order phylogeny

The tree based on gene order is independent from gene sequences and multiple sequence alignments. This tree again shows very good resolution and good overall agreement with single-gene trees (Figure 4.3 B). One important improvement is the finding of the three families *Corynebacteriaceae*, *Mycobacteriacae* and *Nocardiaceae* (*Corynebacterineae*) in one group with strong bootstrap support. The internal branch of *Mycobacteriaceae* containing the six pathogenic species is again more resolved than in the single-gene phylogenies, and the branching mostly agrees with the gene concatenation tree. On the other hand, relationships among the different families of *Frankineae*, *Streptosporangineae* and *Streptomycineae* are less defined here. While all four *Streptomyces* species are places in a single cluster, the three families of *Frankineae* are scattered and their phylogeny unresolved.

Gene content phylogeny

The tree based on gene content (Figure 4.3 C), which uses a third independent source of phylogenetic evidence, is congruent with the 16S and 23S rRNA (Figure 4.2 B and C) trees and with two other whole genome-based trees (Figure 4.3 A and B) for all major branches. *Frankineae*, *Streptomycineae* and *Thermobifida fusca* YX are again clustered in one group. Also, once again *Kineococcus radiotolerans* SRS 30216 is unexpectedly clustered with *Micrococcineae*. The three families *Corynebacteriaceae*, *Mycobacteriacae* and *Nocardiaceae* of the suborder *Corynebacterineae* fall in one monophyletic group. Branching of the six pathogens of *Mycobacteriaceae* is fully resolved and agrees with previous taxonomic classifications. The phylogeny of the three families of the suborder *Micrococcineae* [*Cellulomonadaceae*, *Microbacteriaceae*, *Micrococcaceae*] is completely resolved and branching follows traditional groupings.

Rank order

The rank order of species, using *Streptomyces coelicolor* as our focus species, shows a close relationship between *Streptomycineae*, *Frankineae*, and *Streptosporangineae* (Table 4.3). Among the *Streptomyces* species, *Streptomyces coelicolor* is most closely related to *Streptomyces avermitilis*, followed by *Streptomyces scabies* and *Streptomyces griseus*.

The rank product ranking (Breitling et al., 2004) combines the similarity metrics based on all sources of phylogenetic information (gene sequences, gene order, gene content) into a consensus linear ordering of species. Rankings based on average or median rank yield very similar results [Supporting information: Average and Median rank]. This arrangement can be used, for instance, for arranging species in genome visualization or for determining the weight of species in comparative metabolic modeling.

Consensus tree

As we have shown, all individual tree reconstruction methods show rather low resolution, but good agreement among each other and with classical taxonomical groupings [Supporting information: Phylogenetic trees with bootstrap values]. We also saw that whole genome-based reconstructions generally show a notably improved resolution. Can we use all this information to finally build one, well-supported consensus tree that fully resolves the relationships among our species of interest? To make sure that the final consensus tree contains only branches that are strongly supported by the available evidence, we combined the information from all single-gene and whole genome-based using a majority-rule consensus (Figure 4.4). The resulting tree shows a completely resolved phylogeny of all 45 *Actinomycetales* species, a result that was not achieved independently by any of the individual approaches, whether single gene or genome-based.

The advantage of our integrative approach is most easily seen in the

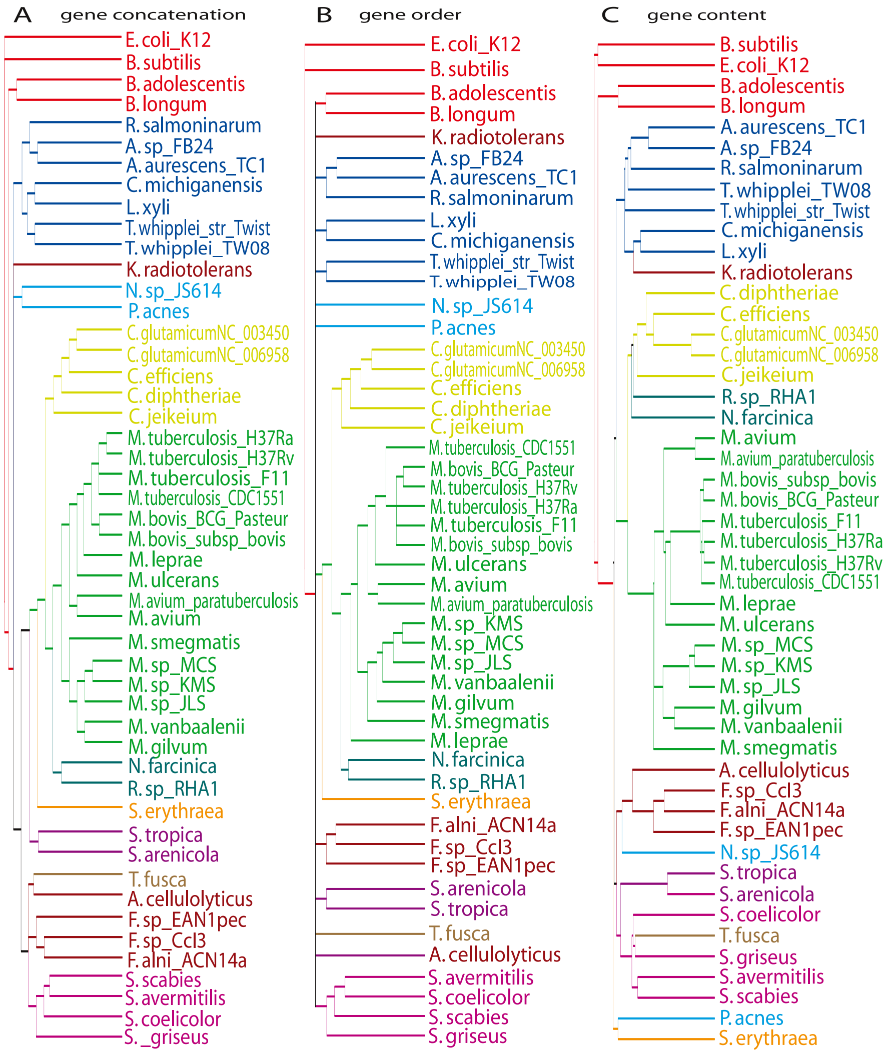

Figure 4.3: *Phylogenetic trees based on whole-genome approaches. (A) Gene concatenation phylogeny based on the concatenation of 155 ubiquitously conserved proteins. (B) Gene order phylogeny based on conserved neighbor relationships along the genome. (C) Gene content phylogeny based on comparison of the set of homologous proteins shared between pairs of species. Coloring as in figure 4.2.*

family *Mycobacteriaceae*: the rRNA-based reconstructions (Figure 4.2), the

Table 4.2: **Consensus ranking of species.***Consensus ranking of species based on their distance from our focus species* Streptomyces coelicolor. *The rank product ranking (Breitling et al., 2004) combines the similarity metrics based on all sources of phylogenetic information (gene sequences, gene order, gene content) into a consensus linear ordering of species. Rankings based on average or median rank yield very similar results (Supplementary information: Average and Median rank). This arrangement can be used, for instance, for arranging species in genome visualization or for determining the weight of species in comparative metabolic modeling.*

Organism	Product rank
Streptomyces coelicolor A3(2)	1
Streptomyces avermitilis MA 4680	2.629032212
Streptomyces scabies strain 8722	2.727747733
Streptomyces griseus strain IFO13350	4.100117402
Thermobifida fusca YX	9.663449262
Nocardioides sp JS614	11.08997499
Acidothermus cellulolyticus 11B	11.39609635
Frankia alni ACN14a	12.06983368
Frankia sp CcI3	12.07899215
Salinispora tropica CNB 440	12.33425514
Kineococcus radiotolerans SRS30216	12.49817983
Frankia sp EAN1pec	13.186375
Salinispora arenicola CNS 205	13.69372661
Saccharopolyspora erythraea NRRL 2338	16.38845893
Nocardia farcinica IFM 10152	17.18458994
Arthrobacter aurescens TC1	18.39612703
Clavibacter michiganensis subsp michiganensis NCPPB 382	18.45680603
Rhodococcus sp RHA1	18.83692805
Propionibacterium acnes KPA171202	19.40448096
Arthrobacter sp FB24	20.27505199
Leifsonia xyli subsp xyli str CTCB07	20.46342192
Renibacterium salmoninarum ATCC 33209	21.4049949
Mycobacterium sp KMS	24.40733018
Mycobacterium sp MCS	24.44682101
Mycobacterium smegmatis str MC2155	24.52484051
Mycobacterium vanbaalenii PYR-1	24.72413092
Mycobacterium sp JLS	25.83612758
Mycobacterium avium subsp paratuberculosis str k10	26.4275277
Mycobacterium gilvum PYR-GCK	26.5103845
Mycobacterium tuberculosis H37Rv	26.7345383
Mycobacterium tuberculosis CDC1551	26.9326884
Mycobacterium tuberculosis H37Ra	27.18827008
Mycobacterium bovis subsp bovis AF2122 97	27.212368
Mycobacterium bovis BCG Pasteur 1173P2	27.93566467
Mycobacterium tuberculosis F11	28.13822088
Tropheryma whipplei TW08 27	28.8506511
Mycobacterium avium 104	28.95281287
Tropheryma whipplei str Twist	29.5717043
Corynebacterium jeikeium K411	29.92280148
Mycobacterium ulcerans Agy99	29.96251954
Mycobacterium leprae TN	30.30757177
Corynebacterium diphtheriae NCTC 13129	30.48200786
Corynebacterium efficiens YS-314	30.66788683
Corynebacterium glutamicum ATCC 13032	33.63478094
Corynebacterium glutamicum ATCC 13032 new	33.67214223
Bifidobacterium longum NCC2705	38.99584717
Bifidobacterium adolescentis ATCC 15703	39.88074983
Bacillus subtilis	48.08436608
Escherichia coli K12	48.36125292

species are only divided into broad clusters with poor resolution of inner branches. In the genome-based approaches the resolution improves notably, but the branching in the different trees is not always consistent (Figure 4.3). In the final consensus tree, the subdivisions of the *Mycobacteriaceae* species are fully resolved (Figure 4.4), despite using a strict majority-rule consensus approach, which includes only the best-supported phylogenetic relationships.

In the final consensus reconstruction (Figure 4.4) the four species of *Streptomyces* are grouped consistently in a single cluster. The closest organism is *Thermobifida fusca* YX (suborder *Streptosporangineae*), in agreement with the rank order analysis (Table 4.3). These five species are shown to be the sister group of the genus *Frankineae*, and closely related to the large cluster containing *Micromonosporineae* and *Corynebacterineae*. The suborders *Pseudonocardineae* and *Corynebacterineae* are phylogenetically very close to each other as seen already in two of the genome-based approaches (gene concatenation and gene order) and in several single gene-based trees (*ribosomal protein S1*, *SecY*, *GTPase*, *DNA topoisomerase*). Remarkable is also the placement of *Kineococcus radiotolerans* SRS 30216: previous classifications placed this species close to *Frankineae* (Lee, 2006; Normand et al., 2007; Garrity et al., 2004), while in the consensus tree in figure 4.4 the species is placed among the *Micrococcineae*, in agreement with the most recent phylogeny of the phylum A*ctinobacteria* (Zhi et al., 2009).

4.4 Conclusions

We aimed to determine a comprehensive, detailed and robust phylogeny of *Streptomyces* and its closest relatives in *Actinomycetales* by using single gene-based and genome-based phylogenies. The results from the different approaches were combined in a consensus tree that shows a completely resolved, well-supported phylogeny of the 45 actinomycetes with high resolution of all inner branches.

Overall, the high-level branchings in our consensus tree agree well with traditional taxonomic subdivisions, grouping together those species which belong taxonomically to the same family or the same suborder. The only major exception is *Kineococcus radiotolerans* SRS 30216, which is placed in *Micro-*

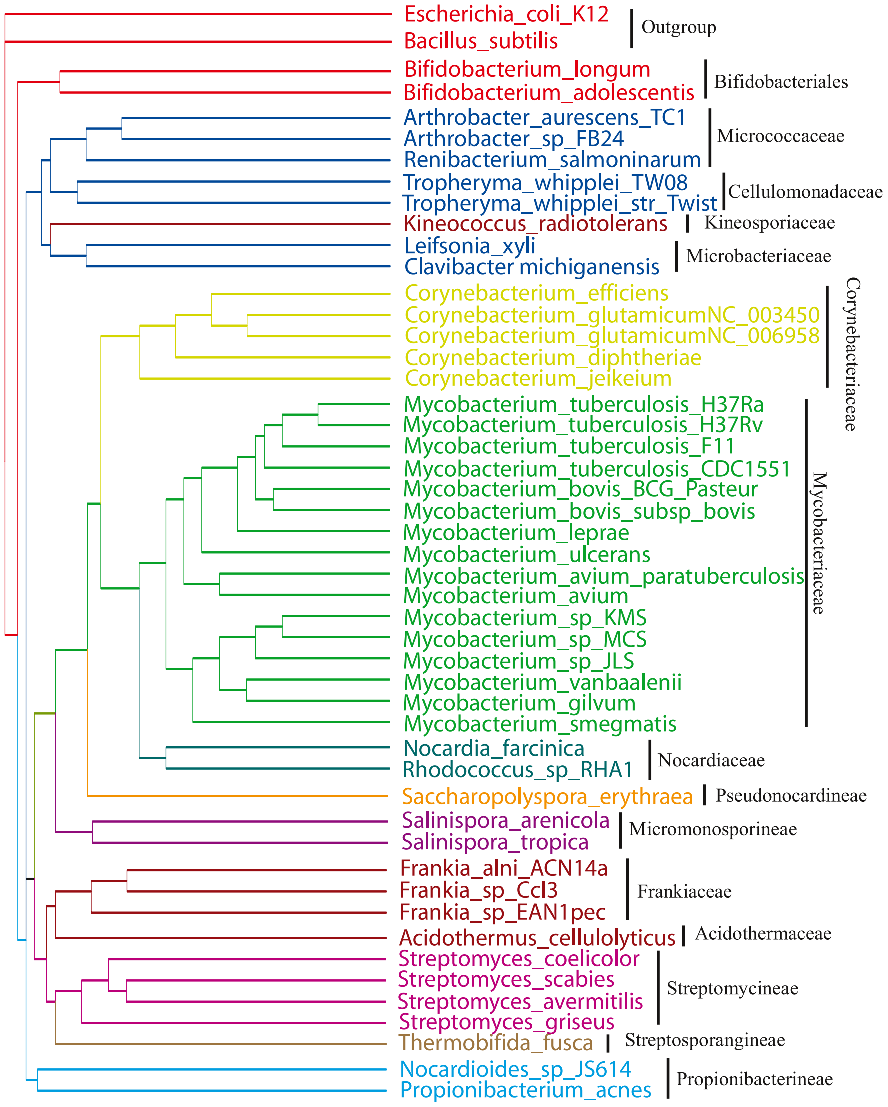

Figure 4.4: *Fully resolved final consensus tree of* Streptomyces *and its relatives, reconstructed from all the independent sources of evolutionary information (gene sequences, gene order, gene content) collected in this study.*

coccineae rather than in *Frankineae* as reported previously (Lilburn and Garrity, 2004; Lee, 2006; Normand et al., 2007). This confirms the general validity of our strategy.

The relationships among the *Streptomyces* species and their relatives are fully resolved only in the final consensus tree (Figure 4.4). The species closest to *Streptomyces* is *Thermobifida fusca* YX, a moderately thermophilic soil bacterium from the suborder *Streptosporangineae*. The *Streptomyces* cluster is furthermore closely related to *Frankineae* species, a group of nitrogen-fixing, nodule-forming bacteria. This finding will be useful for future computational studies exploring the evolution of metabolic capacities in *Streptomyces* species by quantitative genome-based modeling (Price et al., 2004; Breitling et al., 2008).

For the gene concatenation analysis, the reconstruction of phylogenies at the species level could be affected by lateral gene transfer (LGT). Genes that are likely to be subject to LGT should be excluded from the analysis. In our case, the focus on a closely related group of species makes the reliable identification of such LGT candidates impossible. However, the majority of proteins included in the concatenation are encoded by single-copy genes in the majority of genomes, making them unlikely to be subject to widespread LGT. Moreover, we focus on the most highly conserved subset of 155 genes (those present in all species studied), which are least likely to be affected by LGT. This is confirmed by the fact that for all resolved branches the gene concatenation phylogeny and the gene order phylogeny show complete agreement.

In this study we have shown that single-gene analysis can lead to poorly resolved or even misleading phylogenies, and demonstrated a way of using whole-genome analysis to overcome this limitation. We demonstrate that only by a combination of methodologies it is possible to achieve a fully revolved phylogeny with detailed and well-supported inner branches. It will be interesting to apply this method to larger groups of organisms (for instance, all bacteria); the small number of universally conserved protein-coding genes will, however, necessitate modifications in the approach, so that phylogenetic signal from widespread, but non-universal genes can also be included in the analysis.

The phylogenetic reconstructions from the different sources (single gene or whole genome-based phylogenies) agree at the family level with the traditional taxonomy, but at species level they show a much richer, fine-grained picture of the relationships.

The resulting consensus tree, which is supported by a variety of independent sources of evidence (gene sequences, gene order, gene content), will form a solid basis for future comparative genome annotation and across-species modeling of metabolic pathways. This will be an essential resource for the further exploitation of *Streptomyces* species in biotechnology and systems biology research.

4.5 Acknowledgments

M.E.M was funded by a 4 × 4 Ubbo Emmius scholarship from the University of Groningen. E.T. was funded by a Rosalind Franklin Fellowship from the University of Groningen. R.B. is supported by an NWO-Vidi fellowship.

4.6 Supporting information

http://www.sciencedirect.com/science/journal/10557903
Volume 54, Issue 3, March 2010, Pages 763–772

Supplementary material 1: Protein sequence trees

Supplementary material 2: Average and Median rank

Supplementary material 3: List of 155 concatenated proteins

Supplementary material 4: Phylogenetic trees with bootstrap values

Submitted as: **Alam MT**, Medema MH, Takano E, Breitling R (2011): *Comparative genome-scale metabolic modeling of actinomycetes: the topology of essential core metabolism.*

Chapter 5

Comparative genome-scale metabolic modeling of actinomycetes: the topology of essential core metabolism

Abstract

Actinomycetes are very important bacteria, both from a clinical and a commercial perspective. On the one hand, some of them cause severe human and plant diseases; on the other hand, other species are known for their ability to produce antibiotics. This group exhibits a tremendous diversity in terms of ecological niche, as reflected in widely varying genome size.

Here we report the results of a comparative analysis of genome-scale metabolic models of 37 species of Actinomycetales. *Based on single-gene* in silico *knockouts we generated topological and genomic maps of essential enzymes in each organism. We found that low-degree enzymes (i.e., those mostly involved in linear pathways) are generally more essential than high-degree enzymes (i.e., those located at metabolic hubs).*

Combining the collection of genome-wide models, we were able to construct a global enzyme association network and to identify both a conserved "core network" and an "essential core network" of the entire group.

5.1 Introduction

The order *Actinomycetales*, a clade of gram positive bacteria, is one of the best studied groups of bacteria and belongs to the largest phylum of bacterial taxonomy, *Actinobacteria* (Ventura et al., 2007). It is in some respects the most diverse order of bacteria and exhibits biodiversity along several dimensions, including genome length, morphology, ecology, pathogenicity, genomic G+C content, and the number of coding sequences in the genome (Embley and Stackebrandt, 1994; Ventura et al., 2007; Alam et al., 2010a). For instance, genome size varies from a mere 0.9 Mbp in *Tropheryma whipplei* to more than 11.9 Mbp in *Streptomyces bingchenggensis*, which is one of the largest bacterial genomes ever sequenced. Ecologically, some species are soil-dwelling bacteria, some are marine bacteria, others colonize thermal springs (Barabote et al., 2009) or grow on gamma-irradiated surfaces (Phillips et al., 2002) or as plant root symbionts (Normand et al., 2007). The group also contains several important pathogens, affecting animals and plants (Goodfellow and Williams, 1983; Castillo et al., 2002; Tokala et al., 2002): *Mycobacterium tuberculosis* causes tuberculosis (Cole et al., 1998), *Corynebacterium diptheria* infection results in diphtheria (Cerdeño-Tárraga et al., 2003), *Propionibacterium acnes* is the agent of acnes (Leyden, 2001), and *Streptomyces scabies* causes potato scab (Takeuchi et al., 1996). The generally high genomic G+C content varies from 50% in some *Corynebacteria* to over 70% in *Streptomyces* and *Frankia* (Ventura et al., 2007). From a commercial point of view, this group of organisms is also highly important: it has been reported that more than half of the clinically available antibiotics are produced by *Actinomycetales*, and out of these, more than 70% are synthesized by the single genus *Streptomyces* (Goodfellow and Williams, 1983; Castillo et al., 2002; Tokala et al., 2002; Hopwood, 2007).

Despite this amazing diversity, these organisms are clustered together tightly within a single taxonomic branch (Alam et al., 2010a). A number of different studies have examined the genomic commonalities among these organisms (Snel et al., 1999; Sassetti et al., 2001; Yukawa et al., 2007; Alam et al., 2010a). A study to define a set of signature genes has identified a set of 233 conserved proteins which are specific for *Actinobacteria* – the parent phylum of actinomycetes – and which are not shared by other organisms (Gao and Gupta, 2005). These signature proteins were employed to elucidate the interrelationships among different subgroups of *Actinobacteria*. Some additional comparative studies have been performed for specific subsets of actinomycetes (Yukawa et al., 2007; Sevilla et al., 2008; Monot et al., 2009). For instance, in one study, Yukawa and colleagues investigated the core gene set of

Corynebacteria glutamicum and reported 39 novel genes by comparison with other *Corynebacteria* (Yukawa et al., 2007). Most such studies have focused on homology-based profiling for individual genes. No integrative comparative study of similarities and differences at a global scale has been performed to date.

Here we present such a global comparison, based on the genome sequences of more than thirty species of *Actinomycetales*. In the course of this study, we not only performed a comparative study of a large set of actinomycetes, but also established a general pipeline for comparative modeling of large number of species. We can extend this approach to other bacterial groups and include an even large phylogenetic sample in a single study to understand the origin of bacterial metabolic systems.

5.2 Results and Discussion

5.2.1 Model statistics

For the first time, we have reconstructed genome-scale models of 37 actinomycetes and performed comparative analysis of the metabolic systems of these organisms. We used the High-Throughput Genome-scale Metabolic Reconstruction (HTGMR) pipeline of the SEED framework (Henry et al., 2010) to generate preliminary metabolic models, and added a defined minimal set of missing reactions to these models which appeared necessary to simulate biomass production in flux balance analysis in a single common minimal medium condition.

As described earlier, actinomycetes have a very large diversity of genome sizes, ranging from 0.9 Mbp to almost 12 Mbp. Conceivably, this leads to similarly large differences in the size of the genome-scale models, as the number of enzymatic reactions, metabolites, genes and enzymes in the models are all strongly positively correlated with the genome size.

As all preliminarily generated genome-scale metabolic models contain gaps (Feist et al., 2009), we identified a minimal set of essential orphan reactions (gaps that need to be filled) for all preliminary reconstructed mod-

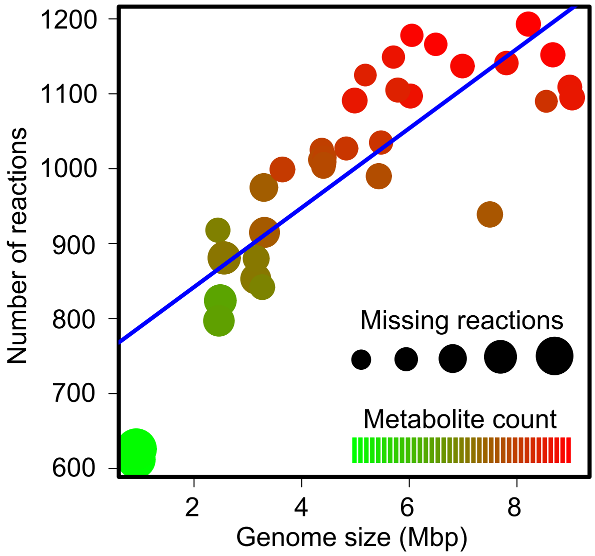

Figure 5.1: ***Model statistics.*** *This figure shows the correlation of various descriptors of the genome-scale models with genome size: total number of reactions, total number of missing reactions and total metabolite count. Larger genomes have more reactions and more of metabolites in their models compared to shorter genomes.*

els according to the need for cellular growth in a defined common minimal medium (see method sections for detailed description). We observed that also the number of such defined missing reactions that need to be added to complete the models is strongly negatively correlated with genome size (Figure 5.1). Plausible causes of this result are that larger genomes generally contain more alternative pathways and duplicate enzymes and that smaller genomes generally encode more enzymes which perform multiple functions (Yus et al., 2009). The models based on these genomes therefore need significantly less gap filling as compared to those based on the shorter genomes Table 5.1).

In total, a set of 283 orphan essential reactions are defined among all models. Of these, 5 reactions constitute gaps in all models: these include N-acetyl-beta-D-mannosaminyltransferase, D-alanine-poly (phosphoribitol) ligase, stearoyl-lipoteichoic acid synthesis, isoheptadecanoyl-lipoteichoic acid

Table 5.1: **Average number of essential orphan reactions added to different suborders of the order *Actinomycetales***

Suborder	gaps	Average genome size (Mb)
Streptomycineae	48	8.7
Micromonosporineae	48	5.5
Pseudonocardineae	51	8.2
Frankineae	53	6.1
Corynebacterineae	54	4.3
Streptosporangineae	54	3.6
Propionibacterineae	73	3.8
Micrococcineae	88	2.3

synthesis and anteisoheptadecanoyl-lipoteichoic acid synthesis. These reactions are necessary for teichoic acid biosynthesis which is particularly poorly characterized in *Actinomycetales* but is considered essential for their biomass production (Rahman et al., 2009).

Other reactions, which are represented by gaps in related subsets of species, are interesting candidates for further experimental investigation: they may be carried out by alternative pathways with uncharacterized biochemistry, or the organisms could have unusual biomass compositions.

5.2.2 *In silico* gene and enzyme knockout to predict essential genes and enzymes

In recent years, several experimental and computational approaches have been used to study the minimal gene set of an organism (Fraser et al., 1995; Mushegian and Koonin, 1996; Sassetti et al., 2001; Kobayashi et al., 2003; Gil et al., 2004; Seringhaus et al., 2006; Gabaldón et al., 2007; Hoffmann et al., 2010). We have found that the genome size of an organism is the strongest determinant of the total number of genes as well as of the total number of predicted essential genes (Figure 5.2 A). The total gene count is positively correlated with genome size (r = 0.92, p-value $< 2.2e{-}16$). On the other hand, the number of predicted essential genes is negatively correlated with gen-

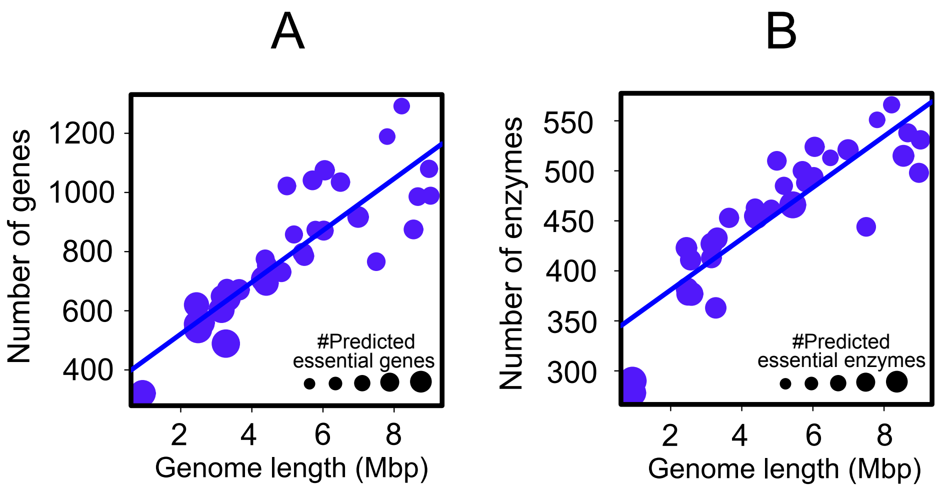

Figure 5.2: ***Correlation of genome length with total number of genes, total number of essential genes, total number of enzymes and total number of essential enzymes.*** ***(A)*** *Positive correlation of genome length (x-axis) with the total number of genes in the metabolic model (r = 0.92, p-value < 2.2e–16), and negative correlation of genome length with the total number of predicted essential genes (r = –0.78, p-value = 8.65e–09), presented by the size of the circles.* ***(B)*** *Positive correlation of genome length (x-axis) with the total number of enzymes in the metabolic model (r = 0.91, p-value = 2.702e–15), and negative correlation of genome length with the total number of predicted essential enzymes (r = –0.64, p-value = 2.046e–05), presented by the size of the circles.*

ome size (r = –0.78, p-value = 8.65e–09) (Figure 5.2 A). Similar results have been reported earlier for gene essentiality of the small genome of *Mycoplasma genitalium*, for which 382 out of 482 genes were reported to be essential (Glass et al., 2006).

Due to the presence of duplicate genes or alternative pathways, the deletion of single genes sometimes has no effect on an organism's phenotype (Gu et al., 2003). If a gene has another functionally equivalent homolog that is expressed under the right conditions, knocking out that particular gene will have no effect in cellular growth because the required enzyme would still be available and the reaction would still be carried out.

In addition to determining essential genes, we also performed enzyme knockouts *in silico* in order to determine a minimal set of essential metabolic enzymes. The total number of enzymes in our models varies from 278 (*Tropheryma whipplei* str Twist) to 566 (*Saccharopolyspora erythraea*), and similar to the total genes counts the total number of enzymes it is also significantly positively correlated to the genome length (r = 0.91, p-value = 2.702e–15). The total number of essential enzymes, like the number of essential genes, negatively correlates with genome size (r = –0.64, p-value = 2.046e–05), but not as strongly, as the existence of a higher number of duplicate genes for essential enzymes in larger genomes boosts the correlation in the latter (Figure 5.2 B).

5.2.3 Pathway distribution of core and predicted essential enzymes

Obviously, it was highly interesting to investigate in which biological pathways the universally conserved enzymes and predicted universally essential enzymes function. We therefore used general KEGG annotations (Kanehisa et al., 2010) to categorize the enzymes into pathway classes. The distribution of universally conserved enzymes among pathways was significantly different after addition of the essentiality criterion (Figure 5.3).

Most notably, core carbohydrate metabolism was quite highly conserved, encompassing a relatively large proportion of the universally conserved genes, but a vast majority of these genes appeared not to be universally essential. Especially surprising was, as can be seen in Supplementary Figures S1 and S2, that genes encoding enzymes for the TCA cycle and glycolysis are quite well conserved, but that none of these enzymes are predicted to be universally essential.

This suggests that the number of alternative pathways in core carbohydrate metabolism (such as the Entner-Doudoroff pathway (Borodina et al., 2005b)) that allow diversion of metabolic fluxes in case of loss of enzyme function in the default pathway is particularly high. This may be caused by evolutionary forces optimizing energetics as well as selecting for mutational robustness in these key processes. As in actinomycetes the lion's share of cellular energy is usually harvested through these pathways, but from many different types of substrates, metabolic functional differentiation may have

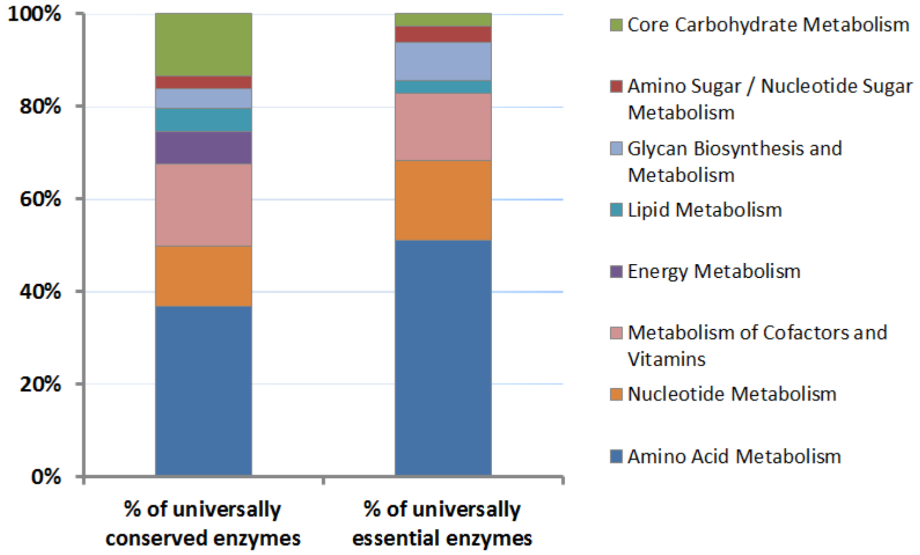

Figure 5.3: ***Pathway distribution of universally conserved enzymes (core enzymes) and universally essential genes (core essential enzymes).*** *The percentages given are based on KEGG pathway categories, and were calculated by locating EC numbers in KEGG pathways using the KEGG API (Kawashima et al., 2003).*

been particularly prominent here during evolution.

5.2.4 Gene essentiality density in genome

We mapped the predicted essential genes to their genome positions and calculated the density of essential genes along the genome (Figure 5.4). Quite strikingly, our result shows that, in most of the organisms, essentiality peaks are away from the center of the genome in contrast to earlier reports (Hopwood, 2006). In some of the genomes, there are two strong essentiality peaks which are located at both extreme ends of the genome (e.g. in *Rhodococcus* sp RHA1), in some cases only one strong peak which is located at any one side of the genome (e.g. *Streptomyces avermitilis*), whereas in relatively few cases the center of the genomes is enriched in essential genes (e.g. *Mycobacterium smegmatis*). One should, however, note that our result is showing only the

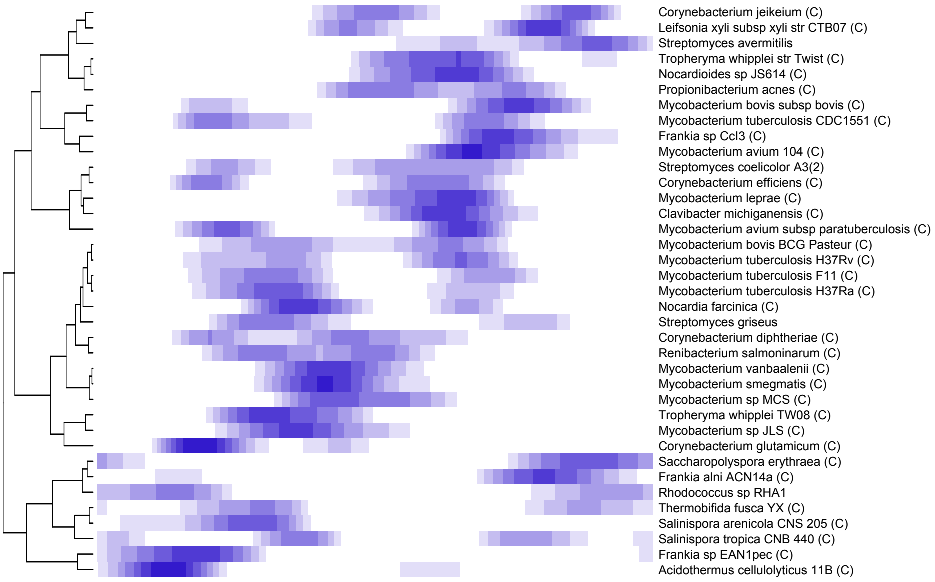

Figure 5.4: ***Genome essentiality map of actinomycetes.*** *This figure shows how the density of essential genes is distributed along the genome in different species. All genomes are scaled to the same total length, and sorted according to the similarity of their profiles. Circular genomes are indicated by a* ***C****.*

location of essential metabolic genes, while other core processes (structural proteins, signalling cascades) may still be enriched more centrally. Moreover, given the facts that the majority of essential enzymes function in amino acid metabolism, nucleotide metabolism and cofactor metabolism, that duplicates of such enzymes often function in secondary metabolism, and that secondary metabolite biosynthesis genes are often located at the ends of actinomycetes chromosomes, a possible explanation for the observed distributions is that in the large genomes with many secondary metabolite biosynthesis gene clusters is a larger number of such enzymes outside the core genome. Possibly, the presence of secondary metabolite biosynthesis genes therefore adds to the mutational robustness of the metabolic network by providing spare genes for highly essential enzymes.

5.2.5 The core metabolic network of *Actinomycetales*

A combined global metabolic network of all 37 organisms in our study is shown in Figure 5.5, where nodes are enzymes and edges are included between nodes that share a common substrate or product (see Methods section for more details). This network is a union of all individual metabolic networks. The conservation of edges across species is shown in a color gradient: red edges are shared by all organisms. Similarly, the conservation of nodes is encoded in the size of node; the large nodes represent enzymes that are universal across species. The color of nodes indicates enzyme essentiality: red enzymes are essential in all species.

One can clearly see from Figure 5.5 that red nodes are larger on average, which means that many universally essential enzymes are universally conserved as well. Out of a total of 753 enzymes, 189 enzymes (25%) are conserved in all organisms ("core enzymes") and 332 enzymes (44%) are conserved in more than 30 organisms (for complete list of enzymes, see Supplementary table S1). Of the 189 universally conserved enzymes, 91 enzymes are predicted to be essential in all studied species of the group ("essential core enzymes"), and 114 enzymes are predicted to be essential in more than 30 organisms.

These are typically enzymes involved in the metabolism of biomass constituents; for instance, amino acid biosynthesis pathways and nucleic acid biosynthesis pathways. Interestingly, 23 enzymes are conserved in all organisms, but predicted to be non-essential in all organisms (large green nodes in Figure 5.5). These enzymes are also mostly involved in synthesis of biomass constituents, but there are apparent alternative pathways in the model. These reactions warrant more detailed study in the future to determine if the supposed alternative routes are indeed available under typical growth conditions.

From the on average 154 essential enzymes per organism, 77% (118 on average) are universally conserved. The fact that on average 23% of the essential enzymes are not conserved suggests that in many cases, different actinomycetes genomes encode different alternative pathways for certain bi-

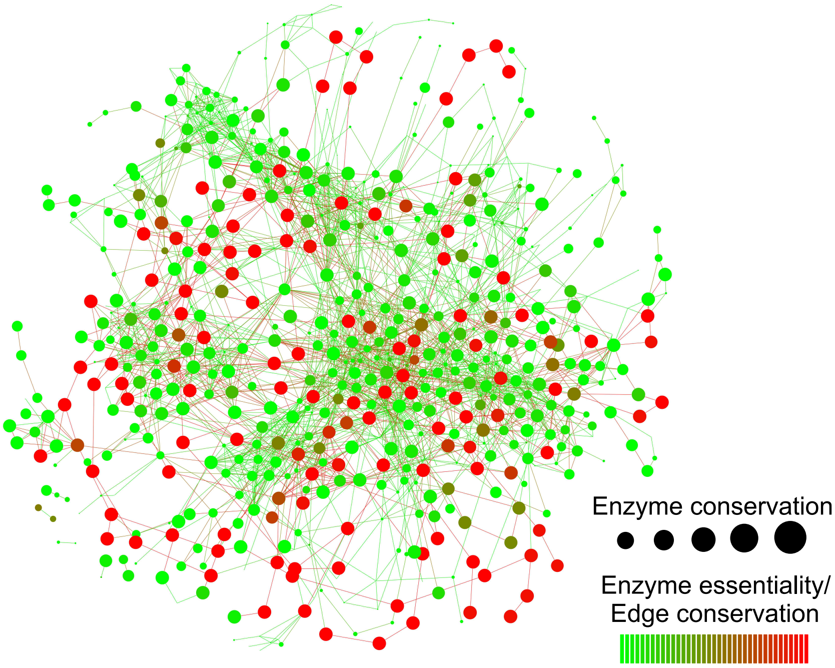

Figure 5.5: ***Global metabolic network of* Actinomycetales**. *A global metabolic network of the entire group of the order* Actinomycetales*, constructed by combining individual species networks, has been shown. Enzyme conservation is presented in the size of the node and a conserved "*core network*" is shown which constitute of all large nodes. Enzyme essentiality has been captured in node color and green color nodes represent universally non-essential and red nodes are showing universally essential enzymes, precisely the "*essential core network*". Enzymes which are essential in some species and non-essential in rest are shown in colors between green and red.*

ological processes.

Out of 189 universally conserved enzymes, 7 are essential in different individual organisms. These enzymes are also particularly interesting for future study, because they imply the loss of specific alternative pathways in selected species (Table 5.2).

Table 5.2: **List of universally conserved enzymes which are predicted to be essential only in single species**

Enzyme EC	Enzyme name	Pathway	Organism
2.4.2.7	AMP:pyrophosphate phosphoribosyltransferase	Purine conversions	*Clavibacter michiganensis*
2.5.1.15	2-Amino-4-hydroxy-6-hydroxymethyl-7,8-dihydropteridine-	Folate biosynthesis	*Corynebacterium diphtheriae*
2.7.1.23	ATP:NAD+ 2′-phosphotransferase	NAD and NADP cofactor biosynthesis	*Leifsonia xyli* subsp. xyli
2.7.4.4	ATP:dTMP phosphotransferase	Pyrimidine metabolism	*Mycobacterium smegmatis*
2.7.6.3	ATP:2-amino-4-hydroxy-6-hydroxymethyl-7,8-dihydropteridine	Folate biosynthesis	*Corynebacterium diphtheriae*
3.6.1.19	XTP pyrophosphohydrolase	Folate biosynthesis	*Mycobacterium smegmatis*
5.3.1.9	D-Glucose-6-phosphate ketol-isomerase	Formaldehyde assimilation: ribulose monophosphate pathway	*Mycobacterium bovis* subsp bovis

5.2.6 Low-degree enzymes are more essential

Microorganisms have several strategies, including gene duplication and presence of alternative pathways, to handle genetic perturbations (Gu et al., 2003). Knocking out individual genes or enzymes may not have an effect on an organism's growth due to the presence of redundant genes or pathways. Network topology, especially the degree distribution of nodes (enzymes) provides significant information about the presumed robustness of microorganisms to perturbation (Hartwell et al., 1999; Barabasi and Oltvai, 2004).

When superimposing our enzyme essentiality data on the degree distribution of nodes, we found that low-degree nodes (in particular nodes of de-

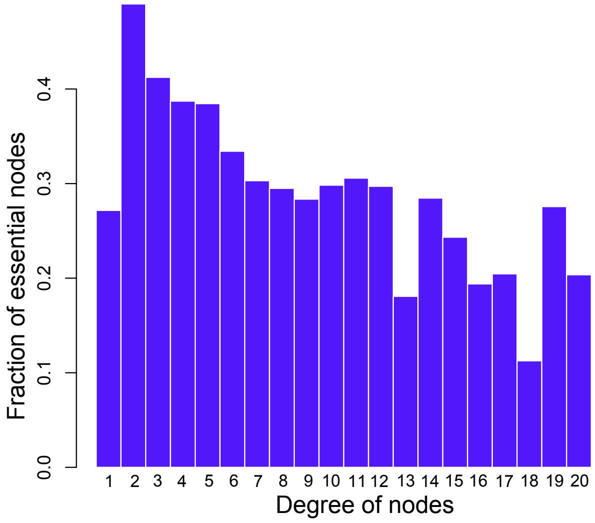

Figure 5.6: ***Essentiality of low degree nodes.*** *This figure shows the fraction of essential nodes for each degree in a metabolic network.*

gree 2) are more likely to be predicted as being essential than high-degree nodes (Figure 5.6). This corroborates the earlier finding at the metabolite level that "low degree metabolites" explain essential reactions in reaction networks of *Escherichia coli*, *Saccharomyces cerevisiae* and *Staphylococcus aureus* (Samal et al., 2006). We could confirm that the low-degree nodes are more essential than high-degree nodes by normalizing the degree essentiality, dividing essential nodes counts with total nodes counts for every degree to see what percentage of nodes are essential for a certain degree (Figure 5.6).

5.2.7 Flux distribution

As described in the previous sections, the 37 genome-scale models differ widely in terms of the number of reactions, number of metabolites, number of genes, and number of enzymes. However, the number of reactions that are active when optimizing for biomass production in a defined common medium is very similar in all organisms (Figure 5.7). Out of these active reactions, on average 52% are carried out by enzymes universally conserved

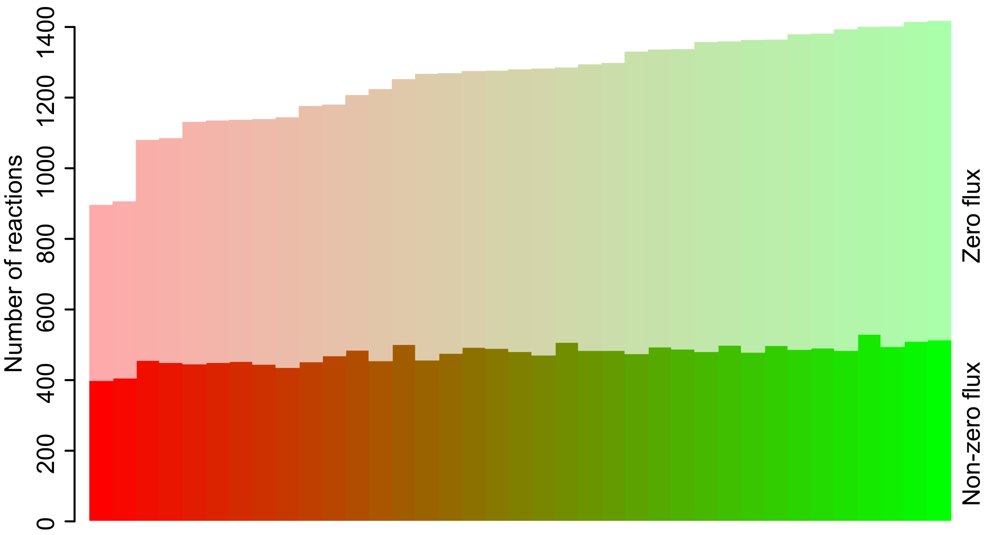

Figure 5.7: ***Distribution of active and inactive reactions depending on model size.*** *The predicted flux profile on a defined common minimal medium indicates the total number of active (non-zero flux) reactions and total number of inactive reactions (zero flux) in each model. It can be seen that the core of active reactions is almost constant, while additional reactions that distinguish organisms are not active under the standardized conditions of our analysis.*

in actinomycetes, and on average 82% of all universally conserved actinomycetes enzymes are active under these conditions. This indicates that despite a large variation in lifestyle and ecological niche, the core metabolic capacities of all actinomycetes are rather similar. Specialization happens outside of central metabolism, e.g. by adding additional assimilation pathways or specific secondary metabolites. Evolutionarily, this hyper-diverse group is therefore characterized by a balance between conservation at the core and innovation at the non-essential fringes of metabolism.

5.2.8 Topological modules and their characteristics

Many biological systems can be partitioned into functional, evolutionary and topological modules (Hartwell et al., 1999; Barabasi and Oltvai, 2004). Modules are defined topologically as densely interconnected groups of vertices,

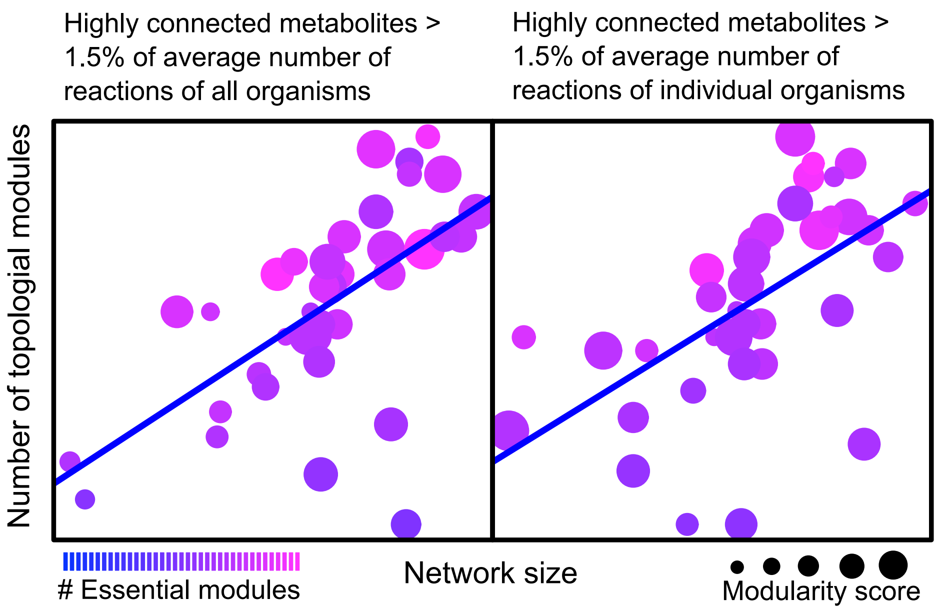

Figure 5.8: ***Modularity score, topological modules, essential topological modules and their relation with network size.*** *The modularity score (Q) is strongly correlated with network size for networks built by removing a fixed number of highly connected metabolites for all organisms, while it is unrelated with network size when the set of highly connected metabolites is defined relative to individual network size. However, total number of modules of a network is determined by the network size. This figure shows the relation of modularity score, total number of modules and total number essential modules of networks by using two different cut-offs for removing the set of highly connected metabolites.*

with only few inter-module links, which usually share a distinct function (Hartwell et al., 1999; Newman, 2006). The modular structure of several biological systems has been studied in detail in the past (Ravasz et al., 2002; Rives and Galitski, 2003; Gagneur et al., 2004; Spirin et al., 2006; Zhao et al., 2007; Kunisawa, 2007).

For all 37 gap-filled analysis-ready genome-scale metabolic models, we inferred enzyme networks after removing some highly connected metabolites, such as ATP (see method section for details of the approach). In order to remove these highly connected metabolites, we used two different cutoffs:

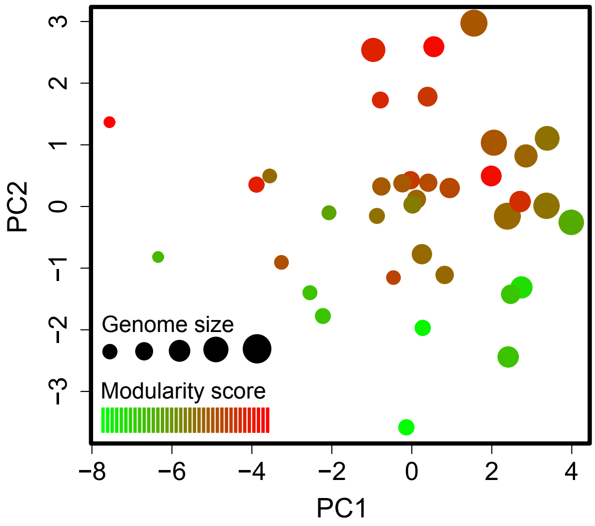

Figure 5.9: ***PCA plot of network topological parameters.***

in one case, we removed all metabolites which are shared by more than 15 reactions, which is 1.5% of the average number of reactions in an average models, in the second case, we removed all metabolites that are shared by >= 1.5% of the total reactions in an individual organism. We used Newman's spectral algorithm (Newman, 2006) to calculate the modularity score (measure of the quality of a division of a network into modules) of each network.

Quite strikingly, both cutoffs resulted in different modularity scores. The modularity of networks built using the first threshold is strongly correlated to network size (r = 0.56, p-value = 0.00027) which is in agreement with Kreimer et al. (Kreimer et al., 2008). On the other hand, modularity in the second case is completely independent of network size (r = –0.047) (Figure 5.8). As the two different cutoffs lead to removal of different numbers of edges from the networks, they result in a different modularity score. However, independent of the cutoff used for removing highly connected metabolites, the number of topological modules strongly correlated with the size of the network (r = 0.62, p-value = 3.571e–05 and r = 0.58, p-value = 0.00018 respectively).

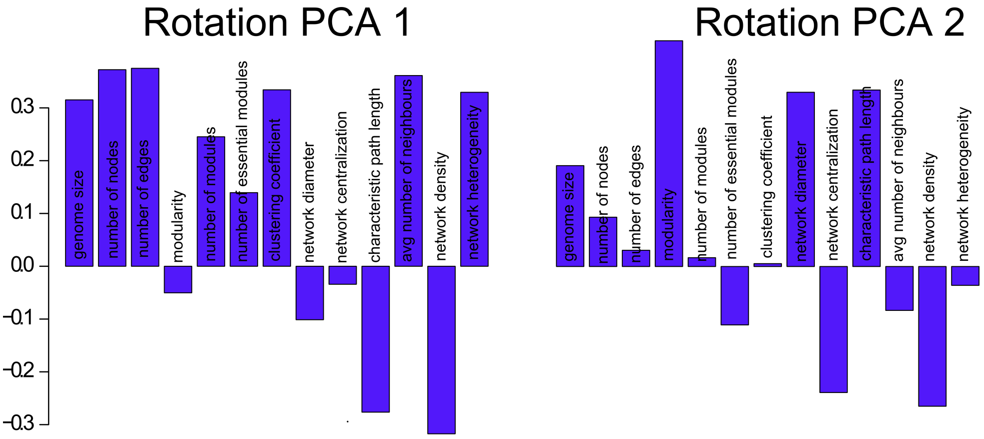

Figure 5.10: ***Rotation of principal component 1 and 2.***

To investigate how many topological modules are predicted to be active in a common minimal medium, we determined the essentiality of nodes of each module of all networks by performing single *in silico* enzyme knockouts. If any enzyme of a module is predicted to be essential, we considered that module as active in the common minimal medium condition. We found that despite having different network sizes and widely varying number of modules, the number of *essential modules* is quite consistent and independent of network size (r = 0.21 and 0.19, respectively).

In addition, we performed a comparative topological analysis of all models: for every network we calculated 13 different network topological parameters and performed a principal component analysis on the varying matrix of parameters (Figure 5.9). Genome size, number of nodes, number of edges, number of modules, clustering coefficient, average number of neighbors and network heterogeneity contribute to the first principal component, whereas the second principal component contains mostly contributions from modularity, network diameter and characteristic path length (Figure 5.10).

5.3 Materials and methods

5.3.1 Genome-scale metabolic model reconstruction

More than fifty genome sequences from the order *Actinomycetales* have already been sequenced, and several individual genome-scale models have been built for several species, including *Corynebacterium glutamicum* (Shinfuku et al., 2009; Kjeldsen and Nielsen, 2009), *Mycobacterium tuberculosis* (Jamshidi and Palsson, 2007; Beste et al., 2007), *Streptomyces coelicolor* (Borodina et al., 2005a; Alam et al., 2010b) and *Streptomyces clavuligerus* (Medema et al., 2010, 2011b). Due to the recent development of the High-Throughput Genome-scale Metabolic Reconstruction (HTGMR) pipeline of the SEED framework (Henry et al., 2010) and several powerful gap filling algorithms (Kharchenko et al., 2004; Green and Karp, 2004; Kumar et al., 2007), it is now possible to reconstruct genome-scale metabolic models and perform comparative modeling of an entire group of organisms. The HTGMR pipeline generates predictive genome-scale models by integrating reaction network annotation and assembly, and thermodynamic analysis for reaction reversibility (Henry et al., 2010). Because genome annotation and model building are done by the same set of tools for all genomes, annotator bias is minimized, which is critical for comparative analysis.

Our analysis included 37 species which had been analyzed earlier in a large genome-based phylogenetic analysis: including 4 *Corynebacteria*, 13 *Mycobacteria*, 3 *Streptomyces*, 3 *Frankia*, 2 *Nocardiaceae*, 1 *Streptosporangineae*, 2 *Micromonosporineae*, 6 *Micrococcineae*, 1 *Pseudonocardineae* and 2 *Propionibacterineae*. We used the HTGMR pipeline to construct a preliminary genome-scale model for each organism and also incorporated an organism-specific biomass reaction as well as a set of exchange reactions.

Since our goal was to study the core metabolic system of the entire group of organisms, we also defined a common biomass reaction and a common minimal medium.

5.3.2 Common minimal medium and defining missing reactions

The universal biomass was defined as listed in supplementary Table S2, based on common biomass constituents of individual organisms. For each individual organism, we collected all minimal media which had been reported in the literature. We then added the minimal set of reactions required to achieve *in silico* growth in the organism-specific minimal media. Finally, based on an organism-specific biomass under rich media conditions, a minimal set of reactions was added to these preliminary models to generate analysis-ready models.

For the comparative analysis, we defined a universal minimal medium, by creating the union of all individual minimal media and iteratively removing compounds to obtain the smallest set of medium components that allows *in silico* growth of all species.

5.3.3 Network construction, division in modules and calculating modularity score

To construct a metabolic network for topological analysis, we translated the stoichiometric matrix into a graph, where each node represents an enzyme and each edge represents a metabolite.

Two nodes of the network are connected if they share a substrate or product (Ma and Zeng, 2003). For each model, some "currency" metabolites which are shared by more than 1.5% of the total number of reactions were removed. Reactions which have one of these metabolites as their only product or substrate were also removed. A common graph for the entire group of organisms was made by combining all individual networks, taking the union of the set of nodes and the set of edges respectively.

5.3.4 Topological analysis

We used a spectral algorithm to calculate the modularity of networks and to divide networks into modules (Newman, 2006). This algorithm constructs the modularity matrix for the network and finds its leading eigenvalue. The corresponding eigenvector then divides the network into two parts according to the signs of the elements; the process is then repeated for each of the part, using the generalized modularity matrix. At any stage, if a proposed split makes a zero or negative contribution to the total modularity, the corresponding subgroup is left undivided.

In addition, the following topological parameters were calculated for each network: clustering coefficient, network diameter, network centralization, characteristic path length, average number of neighbors, network density, and network heterogeneity.

5.4 Conclusions

We have performed genome-scale comparative modeling of the cellular metabolism of 37 genome-sequenced organisms of the order *Actinomycetales*. We find that the characteristic features of the model topologies, including the number of reactions, number of metabolites, number of genes, number of enzymes, number of missing reactions, number of essential genes and number of essential enzymes are highly related to genome size. The predicted metabolic flux profiles indicate that the core of active reactions (non-zero flux) is almost constant, while most additional reactions that differ between organisms are not active under the standardized conditions of our analysis.

Further, we have predicted the list of essential genes and essential enzymes by performing single-gene knockouts and single-enzyme knockouts *in silico* experiments. We mapped the predicted *in silico* gene essentiality data and identified the distribution of essential genes in the genome of each organism.

Additionally, a global metabolic network of entire group of the order *Actinomycetales* was constructed by combining individual species networks, and after superimposing global enzyme essentiality data we determined the key properties of the conserved “core network” and the “essential core network”.

Finally, from genome-scale models we constructed an enzyme association network of individual species, and superimposed predicted enzyme essentiality data on the networks to investigate network topological features. Quite strikingly, we found that the low-degree nodes are more essential than high-degree nodes.

5.5 Acknowledgments

M.T.A was funded by GBB scholarship, University of Groningen. E.T. was funded by a Rosalind Franklin Fellowship from the University of Groningen. R.B. is supported by an NWO-Vidi fellowship.

5.6 Supporting information

Supplementary figure S1: Metabolic core of *Actinomycetales*

Supplementary figure S2: Metabolic essential core of *Actinomycetales*

Supplementary table S1: Complete list of enzymes

Supplementary table S2: Common biomass constituents

Submitted as: **Alam MT**, Takano E, Breitling R (2011) : *Prioritizing orphan proteins for further study using phylogenomics and gene expression profiles in* Streptomyces coelicolor.

Chapter 6

Prioritizing orphan proteins for further study using phylogenomics and gene expression profiles in *Streptomyces coelicolor*

Abstract

Background: Streptomyces coelicolor, *a model organism of antibiotic-producing bacteria, has one of the largest genomes of the bacterial kingdom, including 7825 predicted protein coding genes. Of these genes, a large number, more than 30%, are functionally orphan, i.e. they are encoding hypothetical proteins with unknown function. However, many of these functional orphan genes show interesting gene expression dynamics in large-scale transcriptome analyses.*

Results: *Here we present a new algorithm combining time-course gene expression datasets and comprehensive phylogenomic information to identify a list of high-priority orphan genes, which show the highest level of aggregated evidence of being biologically important. These genes are the ones most generally conserved and showing the most informative expression dynamics along the time course. They often feature conserved neighboring genes as well.*

Conclusions: *The identified high-priority orphan genes are promising candidates to be examined experimentally for further elucidation of their function.*

6.1 Introduction

Here we present an analysis of orphan genes (hypothetical genes with unknown function) in the *Streptomyces coelicolor* genome, combining gene expression analysis and comparative genomics. The aim is to prioritize orphan genes for further study. In our gene expression studies (Nieselt et al., 2010; Alam et al., 2010b), we frequently encountered genes that showed interesting expression patterns, but had no known function. To identify which of these genes merit in-depth experimental analysis, we developed a strategy for prioritizing protein encoding genes for additional characterization, combining phylogenomic information (Alam et al., 2010a) (i.e. the level of evolutionary conservation of each protein), and gene expression data from a large gene expression time series (Nieselt et al., 2010). We postulate that widely conserved proteins that show a physiologically relevant dynamic expression pattern are the most promising candidates for further experimental study, e.g. using gene overexpression and knock-out or knock-down approaches.

The functional annotation of orphan genes is not only relevant for its basic biological interest, but is also an important help for the improvement of genome-scale metabolic models based on genome annotation. These models in their initial form almost always contain gaps that need to be filled by manual curation or automated gap-filling algorithms that add missing essential metabolic activities to the models (Thiele and Palsson, 2010; Henry et al., 2010; Kumar et al., 2007; Alam et al., 2010a; Medema et al., 2010).

During our previous studies of genome-scale metabolic models of *Streptomyces coelicolor* and its relatives, we regularly had to postulate enzymatic functions that had not been assigned to specific proteins in the organisms (Alam et al., 2010a; Medema et al., 2010, 2011b). Assigning specific enzyme-coding genes to these orphan metabolic activities is very important for the subsequent analysis and interpretation of the models, and several approaches have been developed to assign sequences to the orphan metabolic activities: they employ, for example, mRNA co-expression analysis (Kharchenko et al., 2004), phylogenetic profile information (Jothi et al., 2007; Chen and Vitkup, 2006; Snitkin et al., 2006), pattern recognition techniques (Cuff et al., 2009) or comparative genomics (Osterman and Overbeek, 2003). These approaches are organism specific and have mostly been employed for well-studied model organisms such as *Escherichia coli* and *Saccharomyces cerevisiae*.

6.2 Results and Discussion

Of the 7825 predicted protein coding genes in the *Streptomyces coelicolor* genome (Bentley et al., 2002), 2688 (34%) are coding for functionally orphan proteins, i.e. proteins that are annotated as “hypothetical protein”, “conserved protein”, “putative membrane protein” or “putative secreted protein”. Of these orphan proteins, 27 are conserved in all and 384 are present in at least half (22/44) of the 44 analyzed complete actinomycete genomes (see Methods section for a complete species list). 686 orphan proteins are present in at least 11 (25%) and 179 are conserved in at least 33 (75%) actinomycete genomes.

Of the 384 generally conserved actinomycete orphan proteins (i.e., those that are present in at least half of the analyzed genomes), 27 are also encoded in all species in a representative set of five non-actinomycete bacterial genomes (*Bacillus subtilis, Escherichia coli* K12, *Lactobacillus plantarum* WCFS1, *Staphylococcus aureus*, and *Streptococcus pneumonia* AP200), and 73 are present in at least half of the representative bacterial genomes.

Of these 76 ultra-highly conserved bacterial orphan genes, 24 also have putative homologues (reciprocal best BLAST hits) in at least half of the species in a representative set of eight non-bacterial genomes (including the eukaryotes *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Plasmodium falciparum*, *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Homo sapiens*, and the archaea *Haloterrigena turkmenica* and *Methanosarcina acetivorans*). These proteins are therefore almost universally conserved; however, although there seems to be significant conservation of some orphan proteins, none of them is truly universal, i.e. none has a putative homologue in all of the 58 studied genomes. This is most likely due to the fact that some of the included genomes are highly reduced, as a result of the parasitic lifestyle of the organism.

To prioritize the orphan proteins for further characterization, we therefore summarized the phylogenomic information (i.e. the level of evolutionary conservation of each protein) in a single “conservation” score, which expresses the degree of conservation across the three domains examined (acti-

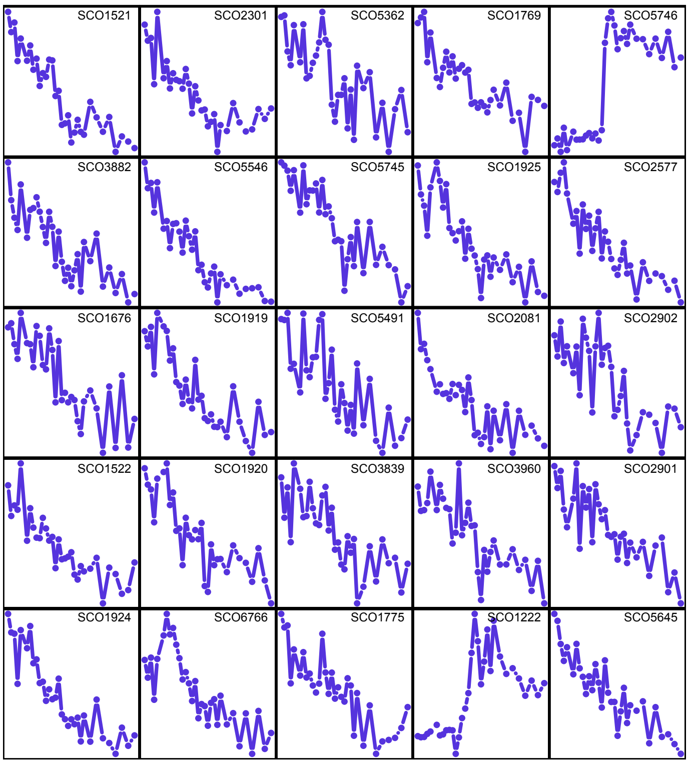

Figure 6.1: *Average expression profile of the top 25 candidate orphan genes.*

nomycetes, bacteria, non-bacteria). This score was combined with a second measure of expression dynamics across a large gene expression time series studying the metabolic switch caused by phosphate starvation. The "expression dynamics" score described in the Methods section identifies genes that show a smooth expression trend across (part of) the time series and favors those genes that show a particularly strong (step-like) expression change at one time point. This is intended to allow to focus on genes that are not only passively following the expression change during nutrient depletion but that show evidence for active regulation, which is indicative of a central function in cellular physiology. Based on the p-value of the "expression dynamics" score, we assigned a rank to each gene, and averaged this value with the rank of the "conservation" score.

Using the averaged conservation and expression dynamics rank, we arrived at a list of 30 top orphan proteins. These were examined in more detail to determine if their function was really unknown: we checked the most recent versions of the Uniprot (The UniProt–Consortium, 2009) and StrepDB database for annotations, performed a PSI-BLAST against the Uniprot database, compared the annotation of the homologs in *E. coli*, yeast and human where these were available, and analyzed the domain architecture using SMART tool (Simple Modular Architecture Research Tool) (Schultz et al., 1998). Using this information, we asked three microbiologist and bioinformaticians to independently score the genes according to their "orphanicity", i.e. their confidence in the absence of a known potential function. The average score of the three raters was combined with the average score of the conservation and expression dynamics to arrive at a final ranking for the most interesting orphan genes for further study: the top genes are those for which we have absolutely no information about their function, that are ultra-highly conserved across species, and show a highly significant dynamics in their gene expression (Table 6.1).

Based on the gene expression profiles (Figure 6.1), the candidate genes SCO5746 and SCO1222 are particularly interesting: they show a very strong switch upon phosphate starvation, and their expression increases upon entry into the stationary phase, similar to the expression pattern of the antibiotic biosynthesis gene clusters *act* and *red*. All other high-priority genes show a decrease of expression along the time course. SCO5746 has a putative uncharacterized homolog in *E. coli* and contains a domain of the DegT/DnrJ-/EryC1/StrS aminotransferase family. The aminotransferase activity was demonstrated for purified StsC protein, which acts as an L-glutamine:scyllo-inosose aminotransferase and catalyses the first amino acid transfer in the biosynthesis of the streptidine subunit of antibiotic streptomycin. It is therefore tempting to speculate that the SCO5746 gene has some role in the biosynthesis of a new antibiotic in *S. coelicolor* as well, and the same might be the case for the completely uncharacterized SCO1222. The closest putative antibiotic biosynthesis clusters are SCO5799-SCO5801 (siderophore synthetase type) and SCO1206-SCO1208 (chalcone synthetase type), both of which seem

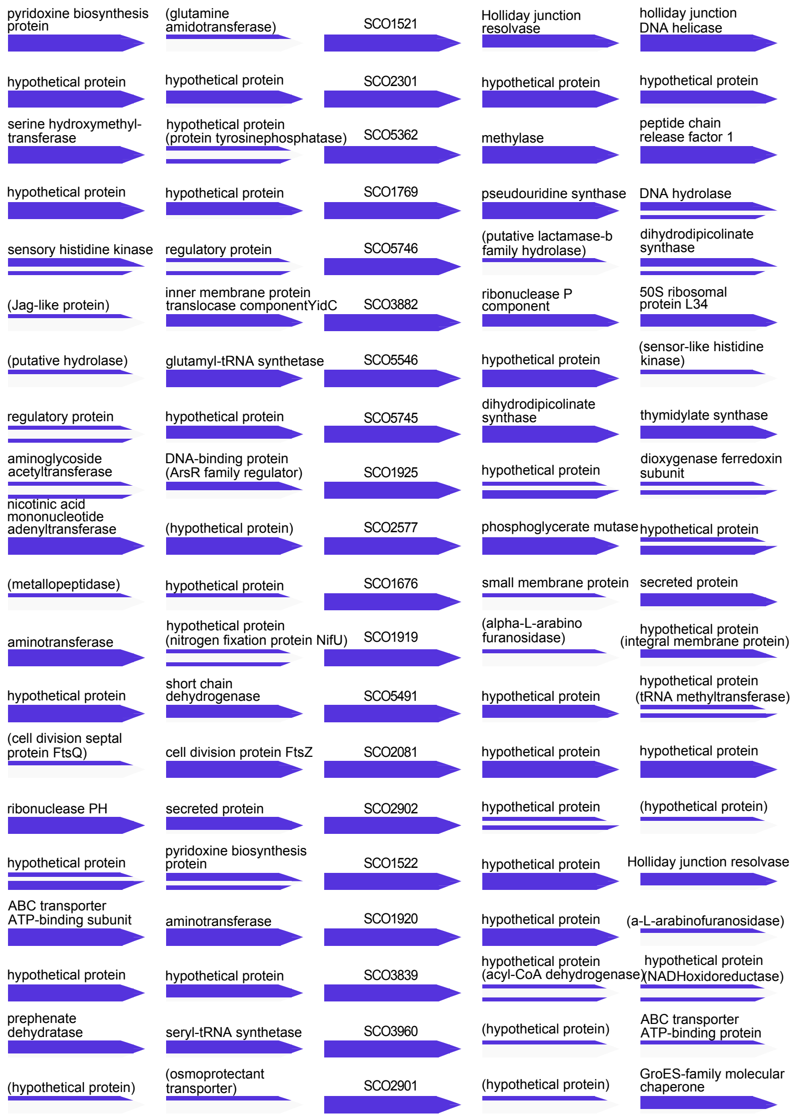
pyridoxine biosynthesis protein
(glutamine amidotransferase)
SCO1521
Holliday junction resolvase
holliday junction DNA helicase
hypothetical protein
hypothetical protein
SCO2301
hypothetical protein
hypothetical protein
serine hydroxymethyl-transferase
hypothetical protein (protein tyrosinephosphatase)
SCO5362
methylase
peptide chain release factor 1
hypothetical protein
hypothetical protein
SCO1769
pseudouridine synthase
DNA hydrolase
sensory histidine kinase
regulatory protein
SCO5746
(putative lactamase-b family hydrolase)
dihydrodipicolinate synthase
(Jag-like protein)
inner membrane protein translocase componentYidC
SCO3882
ribonuclease P component
50S ribosomal protein L34
(putative hydrolase)
glutamyl-tRNA synthetase
SCO5546
hypothetical protein
(sensor-like histidine kinase)
regulatory protein
hypothetical protein
SCO5745
dihydrodipicolinate synthase
thymidylate synthase
aminoglycoside acetyltransferase
DNA-binding protein (ArsR family regulator)
SCO1925
hypothetical protein
dioxygenase ferredoxin subunit
nicotinic acid mononucleotide adenyltransferase
(hypothetical protein)
SCO2577
phosphoglycerate mutase
hypothetical protein
(metallopeptidase)
hypothetical protein
SCO1676
small membrane protein
secreted protein
aminotransferase
hypothetical protein (nitrogen fixation protein NifU)
SCO1919
(alpha-L-arabino furanosidase)
hypothetical protein (integral membrane protein)
hypothetical protein
short chain dehydrogenase
SCO5491
hypothetical protein
hypothetical protein (tRNA methyltransferase)
(cell division septal protein FtsQ)
cell division protein FtsZ
SCO2081
hypothetical protein
hypothetical protein
ribonuclease PH
secreted protein
SCO2902
hypothetical protein
(hypothetical protein)
hypothetical protein
pyridoxine biosynthesis protein
SCO1522
hypothetical protein
Holliday junction resolvase
ABC transporter ATP-binding subunit
aminotransferase
SCO1920
hypothetical protein
(a-L-arabinofuranosidase)
hypothetical protein
hypothetical protein
SCO3839
hypothetical protein (acyl-CoA dehydrogenase)
hypothetical protein (NADHoxidoreductase)
prephenate dehydratase
seryl-tRNA synthetase
SCO3960
(hypothetical protein)
ABC transporter ATP-binding protein
(hypothetical protein)
(osmoprotectant transporter)
SCO2901
(hypothetical protein)
GroES-family molecular chaperone

Figure 6.2: *This figure shows annotation conservation of the neighbors of orphan genes in four sequenced* Streptomyces *genomes. The conserved orphan gene is shown in the centre, and the two neighbors on each side are shown in the form of arrows. Each arrow has four sections, corresponding to the four* Streptomyces *species:* S. coelicolor, S. avermitilis, S. griseus *and* S. scabies. *They are colored in blue where the annotation matches that of* S. coelicolor. *The annotation of the* S. coelicolor *homolog is listed above each gene if it is conserved in at least one of the other species; if at least two of the other species share another annotation, this is listed in brackets.*

Table 6.1: **Top 30 orphan proteins for further study.***The proteins are prioritized according to their conservation across actinomycetes, bacteria and non-bacteria; their expression dynamics (summarized in the p-value); and their orphanicity, i.e. the absence of any functional information.*

Gene Name	Annotation	Final rank	Orphanicity rank	Exp. quantile	p-value	act	bac	non-bac
SCO1521	hypothetical protein	1	1	0.21	3.71E-10	44	5	5
SCO2301	hypothetical protein	6	4	0.34	3.27E-07	43	5	5
SCO5362	hypothetical protein	6.5	9	0.13	2.02E-07	44	4	7
SCO1769	hypothetical protein	8	5	0.12	3.08E-08	40	3	1
SCO5746	hypothetical protein	8	7	0.18	4.38E-18	20	3	1
SCO3882	hypothetical protein	8.5	2	0.18	6.71E-08	38	5	1
SCO5546	hypothetical protein	8.5	14	0.35	7.62E-09	42	3	6
SCO5745	hypothetical protein	9.5	17	0.02	9.49E-10	43	4	6
SCO1925	hypothetical protein	11.5	18	0.09	1.24E-07	44	5	3
SCO2577	hypothetical protein	12	3	0.64	2.66E-07	41	5	1
SCO1676	hypothetical protein	12.5	15	0.32	7.05E-09	31	1	4
SCO1919	hypothetical protein	12.5	11	0.16	5.74E-07	44	4	2
SCO5491	hypothetical protein	12.5	6	0.35	3.07E-07	32	3	3
SCO2081	hypothetical protein	13	8	0.60	2.88E-08	38	2	1
SCO2902	hypothetical protein	14.5	22	0.37	3.05E-07	43	5	4
SCO1522	hypothetical protein	15.5	19	0.19	6.47E-07	43	3	5
SCO1920	hypothetical protein	16	12	0.27	1.71E-06	42	5	5
SCO3839	hypothetical protein	16.5	27	0.35	1.60E-08	35	3	2
SCO3960	hypothetical protein	17.5	13	0.30	5.66E-08	29	5	1
SCO2901	hypothetical protein	18	23	0.36	5.37E-07	41	3	5
SCO1924	hypothetical protein	18.5	20	0.08	6.81E-08	44	1	2
SCO6766	hypothetical protein	18.5	10	0.19	4.55E-08	20	1	2
SCO1775	hypothetical protein	21	16	0.32	3.00E-06	42	4	2
SCO1222	hypothetical protein	22	21	0.43	3.33E-09	27	1	1
SCO5645	hypothetical protein	22	28	0.07	3.11E-07	36	4	2
SCO1530	hypothetical protein	24.5	24	0.03	8.99E-07	43	1	5
SCO2497	hypothetical protein	26.5	29	0.52	2.38E-06	37	5	7
SCO5787	hypothetical protein	27	26	0.12	5.88E-06	44	3	7
SCO2599	hypothetical protein	27.5	25	0.13	4.17E-07	44	1	1
SCO5711	hypothetical protein	29.5	30	0.12	8.65E-06	44	5	5

unlikely candidates for interacting with SCO5746 or SCO1222. However, it is possible that these genes contribute to a dispersed biosynthetic pathway, not involving a dense genomic clustering.

Interestingly, we see a strong neighborhood conservation of most of the candidate orphan genes in other *Streptomyces* species (Figure 6.2). In some cases, the annotation of the neighbors does suggest at least a broad functional category: for example, SCO1521/1522 might be involved in DNA remodeling during recombination, as their conserved neighbors are a Holliday junction resolvase and DNA helicase (RuvABC complex); and SCO2081 might play a role in cell division, matching its conserved neighbor, the cell division protein ftsZ (Jakimowicz et al., 2005). However, most of the conserved neighbors are hypothetical proteins themselves and do not seem to immediately identify a putative function for most of the orphan genes; nonetheless, the neighborhood information will be valuable for the design and interpretation of the most efficient experimental perturbations.

6.3 Materials and methods

6.3.1 Genome sequence analysis

For the phylogenomic profiling, we studied the complete genome sequences of the 44 actinomycete species, that were also used in our earlier phylogenetic study (Alam et al., 2010a): *Arthrobacter aurescens* TC1, *Acidothermus cellulolyticus* 11B, *Bifidobacterium adolescentis* ATCC 15703, *Bifidobacterium longum* NCC2705, *Corynebacterium diphtheriae* NCTC 13129, *Corynebacterium efficiens* YS-314, *Corynebacterium glutamicum* ATCC 13032, *Corynebacterium jeikeium* K411, *Clavibacter michiganensis* subsp michiganensis NCPPB 382, *Frankia alni* ACN14a, *Frankia sp* CcI3, *Frankia sp* EAN1pec, *Kineococcus radiotolerans*, *Leifsonia xyli* subsp xyli str CTCB07, *Mycobacterium avium* subsp, *Paratuberculosis str* k10, *Mycobacterium avium* 104, *Mycobacterium bovis* BCG Pasteur 1173P2, *Mycobacterium bovis* subsp bovis AF2122 97, M*ycobacterium gilvum* PYR-GCK, *Mycobacterium sp* JLS, *Mycobacterium sp* KMS, *Mycobacterium leprae* TN, *Mycobacterium sp* MCS, *Mycobacterium tuberculosis* H37Ra, *Mycobacterium smegmatis* str MC2155, *Mycobacterium tuberculosis* CDC1551, *Mycobacterium tuberculosis* F11, *Mycobacterium tuberculosis* H37Rv, *Mycobacterium ulcerans* Agy99, *Mycobacterium vanbaalenii* PYR-1, *Nocardioides sp* JS614, *Nocardia farcinica* IFM 10152, *Propionibacterium acnes* KPA171202, *Rhodococcus sp* RHA1, *Renibacterium salmoninarum* ATCC 33209, *Salinispora arenicola* CNS 205, *Streptomyces avermitilis* MA 4680, *Saccharopolyspora erythraea* NRRL 2338, *Streptomyces griseus* strain IFO13350, *Streptomyces scabies* strain 8722, *Salinispora tropica* CNB 440, *Thermobifida fusca* YX, *Tropheryma whipplei* str Twist, *Tropheryma whipplei* TW08 27. This was complemented by the genomes of 6 eukaryotes (*Caenorhabditis elegans*, *Arabidopsis thaliana*, *Homo sapiens*, *Plasmodium falciparum* 3D7, *Drosophila melanogaster*, *Saccharomyces cerevisiae*), 2 archaea (*Haloterrigena turkmenica*, *Methanosarcina acetivorans*), and 5 other model bacteria from different taxonomical classes (*Bacillus subtilis*, *Escherichia coli* K12, *Lactobacillus plantarum* WCFS1, *Staphylococcus aureus*, *Streptococcus pneumonia* AP200). Putative homologs were identified as reciprocal best BLAST hits. The conservation score was calculated in three steps: (1) the genes were independently ranked according to the number of species of actinomycetes, other bacteria, and non-bacteria in which they have a putative homolog; (2) their ranks in the bacteria and non-bacteria lists were averaged; and (3) the resulting rank and the rank in the actinomycete list were averaged again to produce the final rank.

6.3.2 Gene expression data

Details about the gene expression dataset and experimental conditions can be found in (Nieselt et al., 2010; Alam et al., 2010b).

6.3.3 Dynamic expression detection

To identify genes that show a dynamic expression along the time course, and in particular genes that have a clear expression switch at one time point, we used the following iterative algorithm (in pseudocode):

```
Data: a vector v of gene expression data
Result: minPvalue, the p-value of the switch-like dynamic expression
minPvalue ← 1;
foreach i in the set (2 to (length(v) - 2)) do
    j ← i + 1;
    MaxWindowSize ← min(i, length (v) - i);
    foreach position p in the set ((i - MaxWindowSize + 1) to i - 1) do
        q ← j + (i - p);
        Pvalue ← p-value of the t-test comparing v[p:i] and v[j:q];
        If (Pvalue < minPvalue) minPvalue ← Pvalue;
    end
end
return minPvalue;
```

Algorithm 1: Algorithm for dynamic expression switch detection

6.4 Conclusions

The aim of this paper was to prioritize protein coding orphan genes (i.e., genes encoding hypothetical proteins with unknown function) for further experimental characterization of their function. We combined two lines of evidence for this purpose: First, we developed an algorithm to score the most interesting dynamic switches in gene expression data. Second, we introduce a conservation score summarzing the level of evolutionary conservation across diverse domains (actinomycetes, other bacteria and non-bacteria). We combined the expression score and the conservation score, and identified a list of 30 high-priority orphan genes, which are promising candidates for future experimental study. In some of the cases, the neighboring genes of the candidate orphan genes show strong conservation and suggest at least a broad functional category for the candidate orphan genes.

6.5 Acknowledgments

M.T.A was funded by GBB scholarship, University of Groningen. E.T. was funded by a Rosalind Franklin Fellowship from the University of Groningen. R.B. is supported by an NWO-Vidi fellowship.

Chapter 7

Conclusions and future perspectives

In the previous chapters, I have presented our studies of a group of actinomycete bacteria whose genome has been sequenced. We have analyzed their metabolic system by constructing genome-scale metabolic models, and by combining different genome-based phylogeny approaches with single gene-based phylogeny approaches we have constructed a fully resolved phylogeny of this group.

In *Streptomyces coelicolor*, a model organism of antibiotic producing bacteria, we investigated the metabolic switch from the exponential phase to the stationary phase during growth: when there are abundant nutrients, bacteria grow exponentially, but as soon as the nutrients get depleted, they switch to stationary phase and often begin to produce secondary metabolites, including some of the commercially important antibiotics. We have investigated the mechanisms involved in the metabolic switch by combining time-scale gene expression data with genome-scale model predictions. Based on discrepancies between model predictions and observed experimental data, we corrected the annotation of some genes and also predicted a list of potential genes which need to be further examined for their role in the synthesis of uncharacterized secondary metabolites.

In another model organism of the antibiotic producing bacteria, *Streptomyces clavuligerus*, we identified key up-regulated antibiotic biosynthetic gene clusters along with some potential up-regulated primary metabolism genes. We constructed a genome-scale metabolic model of this bacterium and optimized growth and antibiotic production, and combined predicted metabolic flux with genome-wide gene expression changes between an industrial clavulanic acid overproduction strain and the parental wild type strain. Up-regulated primary metabolism genes will be useful targets for cell-engineering approaches for overproduction of antibiotics.

We investigated how these commercially interesting bacteria are phylogenetically related to other actinomycetes, some of which are clinically important and cause severe disease in animal and plants. We constructed a fully resolved phylogenetic tree by combining different whole-genome based approaches with the outcome of several single-gene based approaches. Our combinatorial approach to elucidate a robust, comprehensive and completely resolved tree can also be applied to other larger groups of bacteria, and even for all bacteria.

Furthermore, the completely resolved consensus phylogenetic tree of actinomycetes formed the basis for comparative modeling of this hyper-diverse group. We constructed genome-scale metabolic models of a large number of genome-sequenced organisms of the order *Actinomycetales* and identified a conserved "core network" and "essential core network" of the entire group. By using the genome-scale model predictions we were able to identify the commonalities and differences among these actinomycetes on a metabolic systems scale. Additionally, we identified the distribution of essential genes in the genome of each organism by mapping predicted *in silico* gene essentiality data.

Finally, for all functional orphan genes of *Streptomyces coelicolor*, we have applied a dynamic switch detection algorithm on a time–scale gene expression data set, and additional phylogenetic information to identify a list of high priority orphan genes which are promising candidates to be examined in the lab to characterize their function.

In the following sections, I will focus on future perspectives opened up by our research.

7.1 Use and further refinement of actinomycetes genome-scale models

The genome-scale metabolic models presented in this thesis can be used to ask complex questions before setting up experiments in the lab; for instance, which environment is more suitable for bacterial growth, which engineering strategy will be useful for drug targets and antibiotics over production and so on. In this way, the models will be helpful for the identification of new cell-engineering strategies for the synthetic biology of antibiotic production, and in return these strategies will guide the identification of poorly understood aspects of the models.

In the future, when new sets of information, for instance on the functional characterization of orphan genes or on new set of reactions and pathways will become available, this knowledge will be incorporated very easily into the existing models leading to more reliable outcomes of the *in silico* experiments (Akesson et al., 2004; Covert et al., 2008).

7.2 Global integrative models of biological systems

Currently, well-characterized metabolic models for hundreds of microorganisms are being developed at genome scale, but in the vast majority of cases they do not include regulatory effects and intercellular signals. A genome-scale model can only be applied under quasi-steady-state assumptions, and we cannot predict the long-term dynamic behavior of the system, neither can we measure metabolite and enzyme concentrations. Building kinetic models of an entire metabolic system of an organism is usually considered to be infeasible. Similarly, the existing transcription regulatory models of organisms provide tremendous knowledge of the regulatory circuitry, but they typically do not include either signaling information or metabolic information.

The next phase of Systems biology is going to address these points and the first important steps have been taken toward the generation of integrated models (Covert and Palsson, 2002; Akesson et al., 2004; Covert et al., 2008; Shlomi et al., 2007) incorporating regulatory and thermodynamic constraints (Kümmel et al., 2006b,a; Jankowski et al., 2008) and even providing full kinetic parameterization of constraints-based models (Kotte and Heinemann, 2009; Ko et al., 2009; Smallbone et al., 2010).

7.3 Machine learning approaches to predict essential genes

Existing knowledge is used to gain more knowledge; similarly, predictions can be used for further predictions. Genome-scale models have been used quite extensively to predict essential genes and enzymes of microorganisms in defined environmental conditions, and for some of the well curated models, prediction accuracy was more than 90%. To get an idea of the minimal gene set required to sustain a living cell, several experimental and computational approaches have also been applied (Fraser et al., 1995; Mushegian and Koonin, 1996; Gil et al., 2004; Hoffmann et al., 2010; Kobayashi et al., 2003; Gerdes et al., 2003; Sassetti et al., 2001; Glass et al., 2006; Seringhaus et al., 2006; Gabaldón et al., 2007) However, experimental approaches for the estimation of essential genes are expensive and time consuming, while computational approaches require high-throughput data which are not available for every organism. Although there have been some attempts to predict essential genes (Seringhaus et al., 2006; Acencio and Lemke, 2009; Plaimas et al., 2010), these studies were largely based on a very limited number of organisms, which might not be useful in a global prospective.

Now, with the growing number of genome-scale models, we can envisage a new generation of attempts to predict essential genes of microorganisms. Our *in silico* gene essentiality data of of the entire *Actinomycetales* group (presented in chapter 6) can be used to identify characteristic features of gene essentiality, including sequence information, genomic essentiality distribution and network topological information. We can use these characteristic features along with phylogenetic profiles and apply machine learning approaches for sensitive prediction of essential genes.

7.4 Linking genes to orphan metabolic activities

As described in this thesis, all preliminary genome-scale models contain gaps which need to be filled by adding a minimal set of orphan reactions. Assignment of genes to these orphan metabolic activities is a very important step toward completion of the genome-scale model. In our construction of actinomycetes genome-scale models, we found that some orphan reactions were universal, whereas a large number of reactions were organism–specific.

To annotate these orphan reactions and prioritize the potential genes it will in the future be possible to divide the gene networks into topological modules to identify characteristic features of network neighbor of orphan activities.

Following this, by combining characteristic features of network neighboring genes, with gene order and conservation profile we can prioritize the set of genes responsible for orphan activities.

7.5 Functional evolution of metabolic systems

The flux balance analysis approach makes a strong evolutionary assumption about growth optimization. It assumes that microorganisms have evolved in such a way that they will use available nutrients to optimize their growth. It will be a very interesting challenge to explore other objective functions, which might hold for microorganisms that do not achieve maximal growth rates. Amongst the *Actinomycetales*, there are two very important example for such behavior; *Mycobacterium* shows extremely slow growth even in a rich media, and *Streptomyces* sacrifices short-term biomass production for the sake of secondary metabolite biosynthesis, which has more delayed evolutionary benefits. The large-scale comparative modeling introduced in chapter 5 provides a starting point from which to explore how these different evolutionary strategies are reflected in the topology and functional constraints of the metabolic system of different species.

7.6 Application of genome-scale models in Synthetic biology

The goal of Systems Biology is to build predictive computational models which can be used to explore the behavior of biological systems. Once this goal is achieved, Synthetic Biology can use these models for the design and engineering of "de novo" biological systems for specific task. To proceed toward the "de novo" construction of biological circuits, we need to develop models which can not only summarize existing knowledge, but can also incorporate meaningful estimates of missing information. In our study of *Streptomyces* metabolic models, we have explored antibiotic pathways and compared transcriptome data with our predictions. The resulting refined genome-scale models of antibiotic producing organisms will be useful to assist antibiotic over-production in optimal environmental conditions. In particular it will be a major component of ambitious efforts to create synthetic antibiotics production strain, which can overproduce any desired antibiotic compound, of other secondary metabolites of interest. The constraints-based models and their integrated derivatives discussed in the previous section will be indispensable for identifying metabolic bottlenecks and guide the engineering of "pre-conditioned" optimized general overproduction strains (Medema et al., 2011a).

Bibliography

M. Acencio and N. Lemke. Towards the prediction of essential genes by integration of network topology, cellular localization and biological process information. *BMC Bioinformatics*, 10(1):290, 2009.

J. L. Adrio and A. L. Demain. Genetic improvement of processes yielding microbial products. *FEMS Microbiology Reviews*, 30(2):187–214, 2006.

M. Akesson, J. Förster, and J. Nielsen. Integration of gene expression data into genome-scale metabolic models. *Metabolic Engineering*, 6(4):285–293, 2004.

M. T. Alam, M. E. Merlo, E. Takano, and R. Breitling. Genome-based phylogenetic analysis of *Streptomyces* and its relatives. *Molecular Phylogenetics and Evolution*, 54(3):763–772, 2010a.

M. T. Alam, M. E. Merlo, The STREAM Consortium, D. A. Hodgson, E. M. H. Wellington, E. Takano, and R. Breitling. Metabolic modeling and analysis of the metabolic switch in *Streptomyces coelicolor*. *BMC Genomics*, 11:202, 2010b.

D. C. Alexander and S. E. Jensen. Investigation of the *Streptomyces clavuligerus* cephamycin C gene cluster and its regulation by the CcaR protein. *Journal of Bacteriology*, 180(16):4068–4079, 1998.

A. S. Anderson and E. M. Wellington. The taxonomy of *Streptomyces* and related genera. *International Journal of Systematic and Evolutionary Microbiology*, 51(Pt 3):797–814, 2001.

A. K. Apel, A. Sola-Landa, A. Rodríguez-García, and J. F. Martín. Phosphate control of phoA, phoC and phoD gene expression in *Streptomyces coelicolor* reveals significant differences in binding of PhoP to their promoter regions. *Microbiology (Reading, England)*, 153(Pt 10):3527–3537, 2007.

A. P. Arkin. Synthetic cell biology. *Current Opinion in Biotechnology*, 12(6): 638–644, 2001.

A. Barabasi and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nat Rev Genet*, 5(2):101–113, 2004.

R. D. Barabote, G. Xie, D. H. Leu, P. Normand, A. Necsulea, V. Daubin, C. Mdigue, W. S. Adney, X. C. Xu, A. Lapidus, R. E. Parales, C. Detter, P. Pujic, D. Bruce, C. Lavire, J. F. Challacombe, T. S. Brettin, and A. M. Berry. Complete genome of the cellulolytic thermophile *Acidothermus cellulolyticus* 11B provides insights into its ecophysiological and evolutionary adaptations. *Genome Research*, 19(6):1033–1043, 2009.

S. A. Becker, A. M. Feist, M. L. Mo, G. Hannum, B. O. Palsson, and M. J. Herrgard. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA toolbox. *Nature Protocols*, 2(3):727–738, 2007.

S. D. Bentley, K. F. Chater, A. Cerdeño-Tárraga, G. L. Challis, N. R. Thomson, K. D. James, D. E. Harris, M. A. Quail, H. Kieser, D. Harper, A. Bateman, S. Brown, G. Chandra, C. W. Chen, M. Collins, A. Cronin, A. Fraser, A. Goble, J. Hidalgo, T. Hornsby, S. Howarth, C. Huang, T. Kieser, L. Larke, L. Murphy, K. Oliver, S. O'Neil, E. Rabbinowitsch, M. Rajandream, K. Rutherford, S. Rutter, K. Seeger, D. Saunders, S. Sharp, R. Squares, S. Squares, K. Taylor, T. Warren, A. Wietzorrek, J. Woodward, B. G. Barrell, J. Parkhill, and D. A. Hopwood. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature*, 417(6885): 141–147, 2002.

D. J. Beste, T. Hooper, G. Stewart, B. Bonde, C. Avignone-Rossa, M. E. Bushell, P. Wheeler, S. Klamt, A. M. Kierzek, and J. McFadden. GSMN-TB: a web-based genome-scale network model of *Mycobacterium tuberculosis* metabolism. *Genome Biology*, 8(5):R89–R89, 2007.

H. P. Bonarius, G. Schmid, and J. Tramper. Flux analysis of underdetermined metabolic networks: the quest for the missing constraints. *Trends in Biotechnology*, 15(8):308–314, 1997.

I. Borodina, P. Krabben, and J. Nielsen. Genome-scale analysis of *Streptomyces coelicolor* A3(2) metabolism. *Genome Research*, 15(6):820–829, 2005a.

I. Borodina, C. Schller, A. Eliasson, and J. Nielsen. Metabolic network analysis of streptomyces tenebrarius, a streptomyces species with an active entner-doudoroff pathway. *Applied and Environmental Microbiology*, 71(5): 2294–2302, 2005b.

I. Borodina, J. Siebring, J. Zhang, C. P. Smith, G. van Keulen, L. Dijkhuizen, and J. Nielsen. Antibiotic overproduction in *Streptomyces coelicolor* A3 (2) mediated by phosphofructokinase deletion. *The Journal of Biological Chemistry*, 283(37):25186–25199, 2008.

R. Breitling, P. Armengaud, A. Amtmann, and P. Herzyk. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Letters*, 573(1-3):83–92, 2004.

R. Breitling, D. Vitkup, and M. P. Barrett. New surveyor tools for charting microbial metabolic maps. *Nature Reviews. Microbiology*, 6(2):156–161, 2008.

R. N. Brogden, A. Carmine, R. C. Heel, P. A. Morley, T. M. Speight, and G. S. Avery. Amoxycillin/clavulanic acid: a review of its antibacterial activity, pharmacokinetics and therapeutic use. *Drugs*, 22(5):337–362, 1981.

J. R. Brown, C. J. Douady, M. J. Italia, W. E. Marshall, and M. J. Stanhope. Universal trees based on large combined protein sequence data sets. *Nature Genetics*, 28(3):281–285, 2001.

J. Casanova and L. Abel. Genetic dissection of immunity to *Mycobacteria*: the human model. *Annual Review of Immunology*, 20:581–620, 2002.

U. F. Castillo, G. A. Strobel, E. J. Ford, W. M. Hess, H. Porter, J. B. Jensen, H. Albert, R. Robison, M. A. M. Condron, D. B. Teplow, D. Stevens, and D. Yaver. Munumbicins, wide-spectrum antibiotics produced by *Streptomyces* NRRL 30562, endophytic on *Kennedia nigriscans*. *Microbiology (Reading, England)*, 148(Pt 9):2675–2685, 2002.

A. M. Cerdeño-Tárraga, A. Efstratiou, L. G. Dover, M. T. G. Holden, M. Pallen, S. D. Bentley, G. S. Besra, C. Churcher, K. D. James, A. D. Zoysa, T. Chillingworth, A. Cronin, L. Dowd, T. Feltwell, N. Hamlin, S. Holroyd, K. Jagels, S. Moule, M. A. Quail, E. Rabbinowitsch, K. M. Rutherford, N. R. Thomson, L. Unwin, S. Whitehead, B. G. Barrell, and J. Parkhill. The complete genome sequence and analysis of *Corynebacterium diphtheriae* NCTC13129. *Nucleic Acids Research*, 31(22):6516–6523, 2003.

A. M. Cerdeo, M. J. Bibb, and G. L. Challis. Analysis of the prodiginine biosynthesis gene cluster of *Streptomyces coelicolor* A3(2): new mechanisms for chain initiation and termination in modular multienzymes. *Chemistry & Biology*, 8(8):817–829, 2001.

G. L. Challis and D. A. Hopwood. Synergy and contingency as driving forces for the evolution of multiple secondary metabolite production by *Streptomyces* species. *Proceedings of the National Academy of Sciences of the United States of America*, 100 Suppl 2:14555–14561, 2003.

K. F. Chater and G. Chandra. The evolution of development in *Streptomyces* analysed by genome comparisons. *FEMS Microbiology Reviews*, 30(5):651–672, 2006.

K. Chen, Y. Lin, J. Wu, and S. J. Hwang. Enhancement of clavulanic acid production in *Streptomyces clavuligerus* with ornithine feeding. *Enzyme and Microbial Technology*, 32(1):152–156, 2003.

L. Chen and D. Vitkup. Predicting genes for orphan metabolic activities using phylogenetic profiles. *Genome Biology*, 7(2):R17, 2006.

F. D. Ciccarelli, T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork. Toward automatic reconstruction of a highly resolved tree of life. *Science (New York, N.Y.)*, 311(5765):1283–1287, 2006.

S. T. Cole, R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. V. Gordon, K. Eiglmeier, S. Gas, C. E. Barry, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, A. Krogh, J. McLean, S. Moule, L. Murphy, K. Oliver, J. Osborne, M. A. Quail, M. A. Rajandream, J. Rogers, S. Rutter, K. Seeger, J. Skelton, R. Squares, S. Squares, J. E. Sulston, K. Taylor, S. Whitehead, and B. G. Barrell. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, 393(6685):537–544, 1998.

C. Colijn, A. Brandes, J. Zucker, D. S. Lun, B. Weiner, M. R. Farhat, T. Cheng, D. B. Moody, M. Murray, and J. E. Galagan. Interpreting expression data with metabolic flux models: predicting *Mycobacterium tuberculosis* mycolic acid production. *PLoS Computational Biology*, 5(8):e1000489, 2009.

M. W. Covert and B. O. Palsson. Transcriptional regulation in constraints-based metabolic models of *Escherichia coli*. *Journal of Biological Chemistry*, 277(31):28058 –28064, 2002.

M. W. Covert, C. H. Schilling, I. Famili, J. S. Edwards, I. I. Goryanin, E. Selkov, and B. O. Palsson. Metabolic modeling of microbial strains *in silico*. *Trends in Biochemical Sciences*, 26(3):179–186, 2001a.

M. W. Covert, C. H. Schilling, and B. O. Palsson. Regulation of gene expression in flux balance models of metabolism. *Journal of Theoretical Biology*, 213(1):73–88, 2001b.

M. W. Covert, N. Xiao, T. J. Chen, and J. R. Karr. Integrating metabolic, transcriptional regulatory and signal transduction models in *Escherichia coli*. *Bioinformatics*, 24(18):2044 –2050, 2008.

A. L. Cuff, I. Sillitoe, T. Lewis, O. C. Redfern, R. Garratt, J. Thornton, and C. A. Orengo. The CATH classification revisited–architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Research*, 37(Database issue):D310–314, 2009.

V. Daubin, M. Gouy, and G. Perriére. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Research*, 12(7):1080–1090, 2002.

T. Drepper, S. Gross, A. F. Yakunin, P. C. Hallenbeck, B. Masepohl, and W. Klipp. Role of GlnB and GlnK in ammonium control of both nitrogenase systems in the phototrophic bacterium *Rhodobacter capsulatus*. *Microbiology (Reading, England)*, 149(Pt 8):2203–2212, 2003.

M. Durot, P. Bourguignon, and V. Schachter. Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS Microbiology Reviews*, 33(1):164–190, 2009.

S. Egan, P. Wiener, D. Kallifidas, and E. M. Wellington. Phylogeny of *Streptomyces* species and evidence for horizontal transfer of entire and partial antibiotic gene clusters. *Antonie Van Leeuwenhoek*, 79(2):127–133, 2001.

T. M. Embley and E. Stackebrandt. The molecular phylogeny and systematics of the actinomycetes. *Annual Review of Microbiology*, 48:257–289, 1994.

I. Famili, J. Förster, J. Nielsen, and B. O. Palsson. *Saccharomyces cerevisiae* phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proceedings of the National Academy of Sciences of the United States of America*, 100(23):13134 –13139, 2003.

A. M. Feist and B. O. Palsson. The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nat Biotech*, 26(6): 659–667, 2008.

A. M. Feist, M. J. Herrgard, I. Thiele, J. L. Reed, and B. O. Palsson. Reconstruction of biochemical networks in microorganisms. *Nat Rev Micro*, 7(2): 129–143, 2009.

J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.

J. Felsenstein. Phylogeny inference package version 3.67. university of washington, seattle, WA. pp. 981955065, 2007.

W. M. Fitch and E. Margoliash. Construction of phylogenetic trees. *Science (New York, N.Y.)*, 155(760):279–284, 1967.

R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick. Whole-genome random sequencing and assembly of *Haemophilus influenzae* rd. *Science (New York, N.Y.)*, 269(5223):496–512, 1995.

R. M. T. Fleming, I. Thiele, G. Provan, and H. P. Nasheuer. Integrated stoichiometric, thermodynamic and kinetic modelling of steady state metabolism. *Journal of Theoretical Biology*, 264(3):683–692, 2010.

C. M. Fraser, J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton, R. D. Fleischmann, C. J. Bult, A. R. Kerlavage, G. Sutton, J. M. Kelley, R. D. Fritchman, J. F. Weidman, K. V. Small, M. Sandusky, J. Fuhrmann, D. Nguyen, T. R. Utterback, D. M. Saudek, C. A. Phillips, J. M. Merrick, J. F. Tomb, B. A. Dougherty, K. F. Bott, P. C. Hu, T. S. Lucier, S. N. Peterson, H. O. Smith, C. A. Hutchison, and J. C. Venter. The minimal gene complement of *Mycoplasma genitalium*. *Science (New York, N.Y.)*, 270(5235):397–403, 1995.

T. Gabaldón, J. Peretó, F. Montero, R. Gil, A. Latorre, and A. Moya. Structural analyses of a hypothetical minimal metabolism. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 362(1486):1751–1762, 2007.

J. Gagneur, R. Krause, T. Bouwmeester, and G. Casari. Modular decomposition of protein-protein interaction networks. *Genome Biology*, 5(8):R57–R57, 2004.

B. Gao and R. S. Gupta. Conserved indels in protein sequences that are characteristic of the phylum *Actinobacteria*. *International Journal of Systematic and Evolutionary Microbiology*, 55(Pt 6):2401–2412, 2005.

G. M. Garrity, J. A. Bell, and T. G. Lilburn. Taxonomic outline of the prokaryotes. *Bergey's manual of systematic bacteriology, second ed., release 5.0*, 2004.

S. Y. Gerdes, M. D. Scholle, J. W. Campbell, G. Balzsi, E. Ravasz, M. D. Daugherty, A. L. Somera, N. C. Kyrpides, I. Anderson, M. S. Gelfand, A. Bhattacharya, V. Kapatral, M. D'Souza, M. V. Baev, Y. Grechkin, F. Mseeh, M. Y. Fonstein, R. Overbeek, A. Barabsi, Z. N. Oltvai, and A. L. Osterman. Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *Journal of Bacteriology*, 185(19):5673–5684, 2003.

S. Ghorbel, J. Kormanec, A. Artus, and M. Virolle. Transcriptional studies and regulatory interactions between the phoR-phoP operon and the phoU, mtpA, and ppk genes of *Streptomyces lividans* TK24. *Journal of Bacteriology*, 188(2):677–686, 2006.

S. T. F. Gibbon and C. H. House. Whole genome-based phylogenetic analysis of free-living microorganisms. *Nucleic Acids Research*, 27(21):4218 –4222, 1999.

R. Gil, F. J. Silva, J. Pereto, and A. Moya. Determination of the core of a minimal bacterial gene set. *Microbiol. Mol. Biol. Rev.*, 68(3):518–537, 2004.

J. I. Glass, N. Assad-Garcia, N. Alperovich, S. Yooseph, M. R. Lewis, M. Maruf, C. A. Hutchison, H. O. Smith, and J. C. Venter. Essential genes of a minimal bacterium. *Proceedings of the National Academy of Sciences of the United States of America*, 103(2):425 –430, 2006.

A. K. Gombert and J. Nielsen. Mathematical modelling of metabolism. *Current Opinion in Biotechnology*, 11(2):180–186, 2000.

M. Goodfellow and S. T. Williams. Ecology of actinomycetes. *Annual Review of Microbiology*, 37:189–216, 1983.

M. Green and P. Karp. A bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics*, 5(1):76, 2004.

Z. Gu, L. M. Steinmetz, X. Gu, C. Scharfe, R. W. Davis, and W. Li. Role of duplicate genes in genetic robustness against null mutations. *Nature*, 421 (6918):63–66, 2003.

L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray. From molecular to modular cell biology. *Nature*, 1999.

C. S. Henry, M. DeJongh, A. A. Best, P. M. Frybarger, B. Linsay, and R. L. Stevens. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotech*, 28(9):977–982, 2010.

S. R. Henz, D. H. Huson, A. F. Auch, K. Nieselt-Struwe, and S. C. Schuster. Whole-genome prokaryotic phylogeny. *Bioinformatics (Oxford, England)*, 21 (10):2329–2335, 2005.

E. A. Herniou, T. Luque, X. Chen, J. M. Vlak, D. Winstanley, J. S. Cory, and D. R. O'Reilly. Use of whole genome sequence data to infer baculovirus phylogeny. *Journal of Virology*, 75(17):8117–8126, 2001.

C. E. Higgens and R. E. Kastner. *Streptomyces clavuligerus* sp. nov., a beta-Lactam antibiotic producer. *Int J Syst Bacteriol*, 21(4):326–331, 1971.

K. Hoffmann, A. Wollherr, M. Larsen, M. Rachinger, H. Liesegang, A. Ehrenreich, and F. Meinhardt. Facilitation of direct conditional knockout of essential genes in *Bacillus licheniformis* DSM13 by comparative genetic analysis and manipulation of genetic competence. *Appl. Environ. Microbiol.*, 76 (15):5046–5057, 2010.

Z. Hojati, C. Milne, B. Harvey, L. Gordon, M. Borg, F. Flett, B. Wilkinson, P. J. Sidebottom, B. A. M. Rudd, M. A. Hayes, C. P. Smith, and J. Micklefield. Structure, biosynthetic origin, and engineered biosynthesis of calcium-dependent antibiotics from *Streptomyces coelicolor*. *Chemistry & Biology*, 9 (11):1175–1187, 2002.

D. A. Hopwood. Soil to genomics: the *streptomyces* chromosome. *Annual Review of Genetics*, 40(1):1–23, 2006.

D. A. Hopwood. Streptomyces *in Nature and Medicine*. Oxford University Press, 2007.

T. V. Hung, S. Malla, B. C. Park, K. Liou, H. C. Lee, and J. K. Sohng. Enhancement of clavulanic acid by replicative and integrative expression of ccaR and cas2 in *Streptomyces clavuligerus* NRRL3585. *Journal of Microbiology and Biotechnology*, 17(9):1538–1545, 2007.

D. H. Huson and M. Steel. Phylogenetic trees based on gene content. *Bioinformatics (Oxford, England)*, 20(13):2044–2049, 2004.

J. L. Ingraham, O. Maalœ, and F. C. Neidhardt. *Growth of the bacterial cell*. Sinauer Associates, Sinauer, Sunderland, MA, 1983.

D. Jakimowicz, B. Gust, J. Zakrzewska-Czerwinska, and K. F. Chater. Developmental-Stage-Specific assembly of ParB complexes in *Streptomyces coelicolor* hyphae. *Journal of Bacteriology*, 187(10):3572–3580, 2005.

N. Jamshidi and B. O. Palsson. Investigating the metabolic capabilities of *Mycobacterium tuberculosis* H37Rv using the in silico strain iNJ661 and proposing alternative drug targets. *BMC Systems Biology*, 1:26, 2007.

M. D. Jankowski, C. S. Henry, L. J. Broadbelt, and V. Hatzimanikatis. Group contribution method for thermodynamic analysis of complex metabolic networks. *Biophysical Journal*, 95(3):1487–1499, 2008.

H. N. Jnawali, H. C. Lee, and J. K. Sohng. Enhancement of clavulanic acid production by expressing regulatory genes in gap gene deletion mutant of *Streptomyces clavuligerus* NRRL3585. *Journal of Microbiology and Biotechnology*, 20(1):146–152, 2010.

R. Jothi, T. M. Przytycka, and L. Aravind. Discovering functional linkages and uncharacterized cellular pathways using phylogenetic profile comparisons: a comprehensive assessment. *BMC Bioinformatics*, 8:173, 2007.

M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa. KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research*, 38(Database issue):D355–360, 2010.

S. Kawashima, T. Katayama, and M. Kanehisa. KEGG API: a web service using SOAP/WSDL to access the KEGG system. *Genome Informatics*, 14: 673–674, 2003.

P. Kharchenko, D. Vitkup, and G. M. Church. Filling gaps in a metabolic network using expression information. *Bioinformatics (Oxford, England)*, 20 Suppl 1:i178–185, 2004.

H. B. Kim, C. P. Smith, J. Micklefield, and F. Mavituna. Metabolic flux analysis for calcium dependent antibiotic (CDA) production in *Streptomyces coelicolor*. *Metabolic Engineering*, 6(4):313–325, 2004.

K. R. Kjeldsen and J. Nielsen. In silico genome-scale reconstruction and validation of the *Corynebacterium glutamicum* metabolic network. *Biotechnology and Bioengineering*, 102(2):583–597, 2009.

C. Ko, E. Voit, and F. Wang. Estimating parameters for generalized mass action models with connectivity information. *BMC Bioinformatics*, 10(1): 140, 2009.

K. Kobayashi, S. D. Ehrlich, A. Albertini, G. Amati, K. K. Andersen, M. Arnaud, K. Asai, S. Ashikaga, S. Aymerich, P. Bessieres, F. Boland, S. C. Brignell, S. Bron, K. Bunai, J. Chapuis, L. C. Christiansen, A. Danchin, M. Dbarbouill, E. Dervyn, E. Deuerling, K. Devine, S. K. Devine, O. Dreesen, J. Errington, S. Fillinger, S. J. Foster, Y. Fujita, A. Galizzi, R. Gardan, C. Eschevins, T. Fukushima, K. Haga, C. R. Harwood, M. Hecker, D. Hosoya, M. F. Hullo, H. Kakeshita, D. Karamata, Y. Kasahara, F. Kawamura, K. Koga, P. Koski, R. Kuwana, D. Imamura, M. Ishimaru, S. Ishikawa, I. Ishio, D. L. Coq, A. Masson, C. Maul, R. Meima, R. P. Mellado, A. Moir, S. Moriya, E. Nagakawa, H. Nanamiya, S. Nakai, P. Nygaard, M. Ogura, T. Ohanan, M. O'Reilly, M. O'Rourke, Z. Pragai, H. M. Pooley, G. Rapoport, J. P. Rawlins, L. A. Rivas, C. Rivolta, A. Sadaie, Y. Sadaie, M. Sarvas, T. Sato, H. H. Saxild, E. Scanlan, W. Schumann, J. F. M. L. Seegers, J. Sekiguchi, A. Sekowska, S. J. Sror, M. Simon, P. Stragier, R. Studer, H. Takamatsu, T. Tanaka, M. Takeuchi, H. B. Thomaides,

V. Vagner, J. M. van Dijl, K. Watabe, A. Wipat, H. Yamamoto, M. Yamamoto, Y. Yamamoto, K. Yamane, K. Yata, K. Yoshida, H. Yoshikawa, U. Zuber, and N. Ogasawara. Essential *Bacillus subtilis* genes, 2003.

S. Kol, M. E. Merlo, R. A. Scheltema, M. de Vries, R. J. Vonk, N. A. Kikkert, L. Dijkhuizen, R. Breitling, and E. Takano. Metabolomic characterization of the salt stress response in *Streptomyces coelicolor*. *Applied and Environmental Microbiology*, 76(8):2574–2581, 2010.

R. Kolter, D. A. Siegele, and A. Tormo. The stationary phase of the bacterial life cycle. *Annual Review of Microbiology*, 47(1):855–874, 1993.

E. V. Koonin, L. Aravind, and A. S. Kondrashov. The impact of comparative genomics on our understanding of evolution. *Cell*, 101(6):573–576, 2000.

C. Kost, T. Lakatos, I. Bttcher, W. Arendholz, M. Redenbach, and R. Wirth. Non-specific association between filamentous bacteria and fungus-growing ants. *Die Naturwissenschaften*, 94(10):821–828, 2007.

O. Kotte and M. Heinemann. A divide-and-conquer approach to analyze underdetermined biochemical models. *Bioinformatics*, 25(4):519 –525, 2009.

A. Kreimer, E. Borenstein, U. Gophna, and E. Ruppin. The evolution of modularity in bacterial metabolic networks. *Proceedings of the National Academy of Sciences*, 105(19):6976 –6981, 2008.

V. S. Kumar, M. Dasika, and C. Maranas. Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics*, 8(1):212, 2007.

A. Kümmel, S. Panke, and M. Heinemann. Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. *Molecular Systems Biology*, 2:2006.0034–2006.0034, 2006a.

A. Kümmel, S. Panke, and M. Heinemann. Systematic assignment of thermodynamic constraints in metabolic network models. *BMC Bioinformatics*, 7(1):512, 2006b.

T. Kunisawa. Gene arrangements characteristic of the phylum *Actinobacteria*. *Antonie Van Leeuwenhoek*, 92(3):359–365, 2007.

H. A. Lechevalier and M. P. Lechevalier. Biology of actinomycetes. *Annual Review of Microbiology*, 21:71–100, 1967.

J. M. Lee, E. P. Gianchandani, and J. A. Papin. Flux balance analysis in the era of metabolomics. *Briefings in Bioinformatics*, 7(2):140–150, 2006.

J. M. Lee, J. M. Lee, E. P. Gianchandani, J. A. Eddy, and J. A. Papin. Dynamic analysis of integrated signaling, metabolic, and regulatory networks. *PLoS Computational Biology*, 4(5):e1000086, 2008.

S. D. Lee. *Kineococcus marinus* sp. nov., isolated from marine sediment of the coast of jeju, korea. *International Journal of Systematic and Evolutionary Microbiology*, 56(Pt 6):1279–1283, 2006.

J. J. Leyden. The evolving role of *Propionibacterium acnes* in acne. *Seminars in Cutaneous Medicine and Surgery*, 20(3):139–143, 2001.

R. Li and C. A. Townsend. Rational strain improvement for enhanced clavulanic acid production by genetic engineering of the glycolytic pathway in *Streptomyces clavuligerus*. *Metabolic Engineering*, 8(3):240–252, 2006.

T. G. Lilburn and G. M. Garrity. Exploring prokaryotic taxonomy. *International Journal of Systematic and Evolutionary Microbiology*, 54(Pt 1):7–13, 2004.

J. Lin and M. Gerstein. Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. *Genome Research*, 10(6):808–818, 2000.

M. T. López-García, I. Santamarta, and P. Liras. Morphological differentiation and clavulanic acid formation are affected in a *Streptomyces clavuligerus* adpA-deleted mutant. *Microbiology (Reading, England)*, 156(Pt 8):2354–2365, 2010.

H. Ma and A. Zeng. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, 19(2):270 –277, 2003.

R. A. Majewski and M. M. Domach. Simple constrained-optimization view of acetate overflow in *E. coli*. *Biotechnology and Bioengineering*, 35(7):732–738, 1990.

A. Manteca, A. I. Pelaez, R. Zardoya, and J. Sanchez. *Actinobacteria* cyclophilins: phylogenetic relationships and description of new class- and order-specific paralogues. *Journal of Molecular Evolution*, 63(6):719–732, 2006.

T. Marco, D. M. Nathaniel, W. C. Markus, and S. Jorg. Genome-scale metabolic networks. *John Wiley & Sons, Inc. WIREs Syst Biol Med*, 2009.

J. H. Martens, H. Barg, M. J. Warren, and D. Jahn. Microbial production of vitamin B12. *Applied Microbiology and Biotechnology*, 58(3):275–285, 2002.

J. F. Martín and P. Liras. Enzymes involved in penicillin, cephalosporin and cephamycin biosynthesis. *Advances in Biochemical Engineering/Biotechnology*, 39:153–187, 1989.

M. Medema, R. Breitling, R. Bovenberg, and E. Takano. Exploiting plug-and-play synthetic biology for drug discovery and production in microorganisms. *Nature Reviews Microbiology*, 9:131–137, 2011a.

M. H. Medema, A. Trefzer, A. Kovalchuk, M. van den Berg, U. Müller, W. Heijne, L. Wu, M. T. Alam, C. M. Ronning, W. C. Nierman, R. A. L. Bovenberg, R. Breitling, and E. Takano. The sequence of a 1.8-Mb bacterial linear plasmid reveals a rich evolutionary reservoir of secondary metabolic pathways. *Genome Biology and Evolution*, 2:212–224, 2010.

M. H. Medema, M. T. Alam, W. Heijne, M. A. van Berg, U. Mller, A. Trefzer, R. A. L. Bovenberg, R. Breitling, and E. Takano. Genome-wide gene expression changes in an industrial clavulanic acid overproduction strain of *Streptomyces clavuligerus*. *Microbial Biotechnology*, 2011b.

K. Melzoch, M. J. de Mattos, and O. M. Neijssel. Production of actinorhodin by *Streptomyces coelicolor* A3(2) grown in chemostat culture. *Biotechnology and Bioengineering*, 54(6):577–582, 1997.

M. Monot, N. Honore, T. Garnier, N. Zidane, D. Sherafi, A. Paniz-Mondolfi, M. Matsuoka, G. M. Taylor, H. D. Donoghue, A. Bouwman, S. Mays, C. Watson, D. Lockwood, A. Khamispour, Y. Dowlati, S. Jianping, T. H. Rea, L. Vera-Cabrera, M. M. Stefani, S. Banu, M. Macdonald, B. R. Sapkota, J. S. Spencer, J. Thomas, K. Harshman, P. Singh, P. Busso, A. Gattiker, J. Rougemont, P. J. Brennan, and S. T. Cole. Comparative genomic and phylogeographic analysis of *Mycobacterium leprae*. *Nat Genet*, 41(12):1282–1289, 2009.

M. G. Montague and C. A. Hutchison. Gene content phylogeny of herpesviruses. *Proceedings of the National Academy of Sciences of the United States of America*, 97(10):5334–5339, 2000.

A. R. Mushegian and E. V. Koonin. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 93(19):10268–10273, 1996.

M. E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577 –8582, 2006.

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

P. Normand, P. Lapierre, L. S. Tisa, J. P. Gogarten, N. Alloisio, E. Bagnarol, C. A. Bassi, A. M. Berry, D. M. Bickhart, N. Choisne, A. Couloux, B. Cournoyer, S. Cruveiller, V. Daubin, N. Demange, M. P. Francino, E. Goltsman, Y. Huang, O. R. Kopp, L. Labarre, A. Lapidus, C. Lavire,

J. Marechal, M. Martinez, J. E. Mastronunzio, B. C. Mullin, J. Niemann, P. Pujic, T. Rawnsley, Z. Rouy, C. Schenowitz, A. Sellstedt, F. Tavares, J. P. Tomkins, D. Vallenet, C. Valverde, L. G. Wall, Y. Wang, C. Medigue, and D. R. Benson. Genome characteristics of facultatively symbiotic *Frankia* sp. strains reflect host range and host plant biogeography. *Genome Research*, 17 (1):7–15, 2007.

M. A. Oberhardt, B. O. Palsson, and J. A. Papin. Applications of genome-scale metabolic reconstructions. *Mol Syst Biol*, 5, 2009.

Y. Oh, B. O. Palsson, S. M. Park, C. H. Schilling, and R. Mahadevan. Genome-scale reconstruction of metabolic network in *Bacillus subtilis* based on high-throughput phenotyping and gene essentiality data. *The Journal of Biological Chemistry*, 282(39):28791–28799, 2007.

G. J. Olsen and C. R. Woese. Ribosomal RNA: a key to phylogeny. *The FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology*, 7(1):113–123, 1993.

A. Osterman and R. Overbeek. Missing genes in metabolic pathways: a comparative genomics approach. *Current Opinion in Chemical Biology*, 7(2):238–251, 2003.

B. O. Palsson. *Systems Biology Properties of Reconstructed Networks*. Cambridge University Press, University of California, San Diego, 2006.

E. T. Papoutsakis and C. L. Meyer. Equations and calculations of product yields and preferred pathways for butanediol and mixed-acid fermentations. *Biotechnology and Bioengineering*, 27(1):50–66, 1985.

A. S. Paradkar, K. A. Aidoo, and S. E. Jensen. A pathway-specific transcriptional activator regulates late steps of clavulanic acid biosynthesis in *Streptomyces clavuligerus*. *Molecular Microbiology*, 27(4):831–843, 1998.

R. W. Phillips, J. Wiegel, C. J. Berry, C. Fliermans, A. D. Peacock, D. C. White, and L. J. Shimkets. *Kineococcus radiotolerans* sp. nov., a radiation-resistant, gram-positive bacterium. *International Journal of Systematic and Evolutionary Microbiology*, 52(Pt 3):933–938, 2002.

K. Plaimas, R. Eils, and R. König. Identifying essential genes in bacterial metabolic networks with machine learning methods. *BMC Systems Biology*, 4:56–56, 2010.

N. D. Price, J. A. Papin, C. H. Schilling, and B. O. Palsson. Genome-scale microbial in silico models: the constraints-based approach. *Trends in Biotechnology*, 21(4):162–169, 2003.

N. D. Price, J. L. Reed, and B. O. Palsson. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nature Reviews. Microbiology*, 2(11):886–897, 2004.

O. Rahman, L. G. Dover, and I. C. Sutcliffe. Lipoteichoic acid biosynthesis: two steps forwards, one step sideways? *Trends in Microbiology*, 17(6):219–225, 2009.

E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. Barabasi. Hierarchical organization of modularity in metabolic networks. *Science*, 297 (5586):1551–1555, 2002.

J. L. Reed, T. D. Vo, C. H. Schilling, and B. O. Palsson. An expanded genome-scale model of *Escherichia coli* k-12 (iJR904 GSM/GPR). *Genome Biology*, 4 (9):R54, 2003.

A. W. Rives and T. Galitski. Modular organization of cellular networks. *Proceedings of the National Academy of Sciences of the United States of America*, 100(3):1128 –1133, 2003.

A. Rodríguez-García, A. de la Fuente, R. Pérez-Redondo, J. F. Martín, and P. Liras. Characterization and expression of the arginine biosynthesis gene cluster of *Streptomyces clavuligerus*. *Journal of Molecular Microbiology and Biotechnology*, 2(4):543–550, 2000.

Rokas and Holland. Rare genomic changes as a tool for phylogenetics. *Trends in Ecology & Evolution (Personal Edition)*, 15(11):454–459, 2000.

J. Romero, P. Liras, and J. F. Martín. Dissociation of cephamycin and clavulanic acid biosynthesis in *Streptomyces clavuligerus*. *Applied Microbiology and Biotechnology*, 20(5):318–325, 1984.

D. B. Roszak and R. R. Colwell. Survival strategies of bacteria in the natural environment. *Microbiological Reviews*, 51(3):365–379, 1987.

N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987.

S. P. Salowe, E. N. Marsh, and C. A. Townsend. Purification and characterization of clavaminate synthase from *Streptomyces clavuligerus*: an unusual oxidative enzyme in natural product biosynthesis. *Biochemistry*, 29(27):6499–6508, 1990.

A. Samal, S. Singh, V. Giri, S. Krishna, N. Raghuram, and S. Jain. Low degree metabolites explain essential reactions and enhance modularity in biological networks. *BMC Bioinformatics*, 7(1):118, 2006.

C. M. Sassetti, D. H. Boyd, and E. J. Rubin. Comprehensive identification of conditionally essential genes in mycobacteria. *Proceedings of the National Academy of Sciences of the United States of America*, 98(22):12712 –12717, 2001.

P. S. Saudagar, S. A. Survase, and R. S. Singhal. Clavulanic acid: a review. *Biotechnology Advances*, 26(4):335–351, 2008.

J. Schultz, F. Milpetz, P. Bork, and C. P. Ponting. SMART, a simple modular architecture research tool: Identification of signaling domains. *Proceedings of the National Academy of Sciences of the United States of America*, 95(11):5857 –5864, 1998.

M. Seringhaus, A. Paccanaro, A. Borneman, M. Snyder, and M. Gerstein. Predicting essential genes in fungal genomes. *Genome Research*, 16(9):1126–1135, 2006.

I. Sevilla, L. Li, A. Amonsin, J. Garrido, M. Geijo, V. Kapur, and R. Juste. Comparative analysis of *Mycobacterium avium* subsp. paratuberculosis isolates from cattle, sheep and goats by short sequence repeat and pulsed-field gel electrophoresis typing. *BMC Microbiology*, 8(1):204, 2008.

N. Shahab, F. Flett, S. G. Oliver, and P. R. Butler. Growth rate control of protein and nucleic acid content in *Streptomyces coelicolor* A3(2) and *Escherichia coli* B/r. *Microbiology (Reading, England)*, 142 (Pt 8):1927–1935, 1996.

Y. Shinfuku, N. Sorpitiporn, M. Sono, C. Furusawa, T. Hirasawa, and H. Shimizu. Development and experimental verification of a genome-scale metabolic model for *Corynebacterium glutamicum*. *Microbial Cell Factories*, 8 (1):43, 2009.

T. Shlomi, Y. Eisenberg, R. Sharan, and E. Ruppin. A genome-scale computational study of the interplay between transcriptional regulation and metabolism. *Molecular Systems Biology*, 3:101–101, 2007.

K. Smallbone, E. Simeonidis, N. Swainston, and P. Mendes. Towards a genome-scale kinetic model of cellular metabolism. *BMC Systems Biology*, 4(1):6, 2010.

B. Snel, P. Bork, and M. A. Huynen. Genome phylogeny based on gene content. *Nature Genetics*, 21(1):108–110, 1999.

B. Snel, M. A. Huynen, and B. E. Dutilh. Genome trees and the nature of genome evolution. *Annual Review of Microbiology*, 59:191–209, 2005.

E. S. Snitkin, A. M. Gustafson, J. Mellor, J. Wu, and C. DeLisi. Comparative assessment of performance and genome dependence among phylogenetic profiling methods. *BMC Bioinformatics*, 7:420, 2006.

J. Sourdis and M. Nei. Relative efficiencies of the maximum parsimony and distance-matrix methods in obtaining the correct phylogenetic tree. *Molecular Biology and Evolution*, 5(3):298–311, 1988.

V. Spirin, M. S. Gelfand, A. A. Mironov, and L. A. Mirny. A metabolic network in the evolutionary context: multiscale structure and modularity. *Proceedings of the National Academy of Sciences of the United States of America*, 103(23):8774–8779, 2006.

E. Stackebrandt, F. A. Rainey, and N. L. Ward-Rainey. Proposal for a new hierarchic classification system, *Actinobacteria* classis nov. *Int J Syst Bacteriol*, 47(2):479–491, 1997.

K. Tahlan, C. Anders, and S. E. Jensen. The paralogous pairs of genes involved in clavulanic acid and clavam metabolite biosynthesis are differently regulated in *Streptomyces clavuligerus*. *Journal of Bacteriology*, 186(18): 6286–6297, 2004.

K. Tahlan, C. Anders, A. Wong, R. H. Mosher, P. H. Beatty, M. J. Brumlik, A. Griffin, C. Hughes, J. Griffin, B. Barton, and S. E. Jensen. 5S clavam biosynthetic genes are located in both the clavam and paralog gene clusters in *Streptomyces clavuligerus*. *Chemistry & Biology*, 14(2):131–142, 2007.

T. Takeuchi, H. Sawada, F. Tanaka, and I. Matsuda. Phylogenetic analysis of *Streptomyces* spp. causing potato scab based on 16S rRNA sequences. *International Journal of Systematic Bacteriology*, 46(2):476–479, 1996.

The UniProt–Consortium. The universal protein resource (UniProt) in 2010. *Nucleic Acids Research*, 38(Database):D142–D148, 2009.

I. Thiele and B. O. Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protocols*, 5(1):93–121, 2010.

J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22(22):4673–4680, 1994.

R. K. Tokala, J. L. Strap, C. M. Jung, D. L. Crawford, M. H. Salove, L. A. Deobald, J. F. Bailey, and M. J. Morra. Novel plant-microbe rhizosphere interaction involving *Streptomyces lydicus* WYEC108 and the pea plant *Pisum sativum*. *Applied and Environmental Microbiology*, 68(5):2161–2171, 2002.

A. Varma and B. O. Palsson. Metabolic flux balancing: basic concepts, scientific and practical use. *Nat Biotech*, 12(10):994–998, 1994.

A. Varma, B. W. Boesch, and B. O. Palsson. Stoichiometric interpretation of *Escherichia coli* glucose catabolism under various oxygenation rates. *Applied and Environmental Microbiology*, 59(8):2465–2473, 1993.

M. Ventura, C. Canchaya, A. Tauch, G. Chandra, G. F. Fitzgerald, K. F. Chater, and D. van Sinderen. Genomics of *Actinobacteria*: tracing the evolutionary history of an ancient phylum. *Microbiology and Molecular Biology Reviews: MMBR*, 71(3):495–548, 2007.

P. H. Viollier, W. Minas, G. E. Dale, M. Folcher, and C. J. Thompson. Role of acid metabolism in *Streptomyces coelicolor* morphological differentiation and antibiotic biosynthesis. *Journal of Bacteriology*, 183(10):3184–3192, 2001a.

P. H. Viollier, K. T. Nguyen, W. Minas, M. Folcher, G. E. Dale, and C. J. Thompson. Roles of aconitase in growth, metabolism, and morphological differentiation of *Streptomyces coelicolor*. *Journal of Bacteriology*, 183(10): 3193–3203, 2001b.

A. Wagner. Energy constraints on the evolution of gene expression. *Molecular Biology and Evolution*, 22(6):1365–1374, 2005.

A. Wagner. Energy costs constrain the evolution of gene expression. *Journal of Experimental Zoology. Part B, Molecular and Developmental Evolution*, 308 (3):322–324, 2007.

C. R. Woese. Bacterial evolution. *Microbiological Reviews*, 51(2):221–271, 1987.

K. Yanai, T. Murakami, and M. Bibb. Amplification of the entire kanamycin biosynthetic gene cluster during empirical strain improvement of *Streptomyces kanamyceticus*. *Proceedings of the National Academy of Sciences of the United States of America*, 103(25):9661–9666, 2006.

H. Yukawa, C. A. Omumasaba, H. Nonaka, P. Kos, N. Okai, N. Suzuki, M. Suda, Y. Tsuge, J. Watanabe, Y. Ikeda, A. A. Vertes, and M. Inui. Comparative analysis of the *Corynebacterium glutamicum* group and complete genome sequence of strain r. *Microbiology*, 153(4):1042–1058, 2007.

E. Yus, T. Maier, K. Michalodimitrakis, V. van Noort, T. Yamada, W. Chen, J. A. H. Wodke, M. Gell, S. Martnez, R. Bourgeois, S. Khner, E. Raineri, I. Letunic, O. V. Kalinina, M. Rode, R. Herrmann, R. Gutirrez-Gallego, R. B. Russell, A. Gavin, P. Bork, and L. Serrano. Impact of genome reduction on bacterial metabolism and its regulation. *Science (New York, N.Y.)*, 326(5957): 1263–1268, 2009.

H. Zhang, W. Zhang, Y. Jin, M. Jin, and X. Yu. A comparative study on the phylogenetic diversity of culturable actinobacteria isolated from five marine sponge species. *Antonie Van Leeuwenhoek*, 93(3):241–248, 2008.

J. Zhao, G. Ding, L. Tao, H. Yu, Z. Yu, J. Luo, Z. Cao, and Y. Li. Modular co-evolution of metabolic networks. *BMC Bioinformatics*, 8(1):311, 2007.

X. Zhi, W. Li, and E. Stackebrandt. An update of the structure and 16S rRNA gene sequence-based definition of higher ranks of the class *Actinobacteria*, with the proposal of two new suborders and four new families and emended descriptions of the existing higher taxa. *International Journal of Systematic and Evolutionary Microbiology*, 59(Pt 3):589–608, 2009.

M. C. Zuneda, J. J. Guillenea, J. B. Dominguez, A. Prado, and F. M. Goni. Lipid composition and protoplast-forming capacity of *Streptomyces antibioticus*. *Lipids*, 19(3):223–228, 1984.

Dutch summary

Streptomyces-bacteriën worden vaak "antibiotica-fabriekjes" genoemd, vanwege het feit dat ze een groot aantal klinisch belangrijke chemische stoffen kunnen produceren. Ze behoren tot de orde van de Actinomycetales, een groep organismen die sterke biologische diversiteit vertoont in genoomgrootte, pathogeniciteit, ecologische niches en hun vermogens om secondaire metabolieten te produceren. Dit proefschrift begint met een inleiding over het geslacht Streptomyces en zijn familie, alsmede de beschrijving van modelleringstechnieken die gebruikt worden voor het analyseren van hun metabole functies (**Hoofdstuk 1**). We hebben het metabole system van twee antibiotica-producerende modelorganismen, *Streptomyces coelicolor* (**Hoofdstuk 2**) en *Streptomyces clavuligerus* (**Hoofdstuk 3**), verkend, en met computermodellen de mechanismen achter de antibiotica-productie onderzocht. Om te begrijpen hoe deze antibiotica-producerende soorten zich fylogenetisch gezien verhouden tot andere soorten onder de actinomyceten, hebben we hun fylogenie tot in detail ontrafeld en een algemeen bruikbare robuuste methode ontwikkeld om eenduidige en consistente fylogenetische bomen te construeren uit genoomsequenties (**Hoofdstuk 4**). De resultaten van de fylogenetische studie vormden de basis voor het grootschalig metabolisch modelleren van verschillende soorten, met behulp waarvan we zowel metabolische als topologische overeenkomsten en verschillen identificeerden tussen de actinomyceten (**Hoofdstuk 5**). Vervolgens hebben we, door fylogenetische informatie met gen-expressiedata te combineren, de "orphan"-genen van *Streptomyces coelicolor* geïdentificeerd die het meest veelbelovend zijn voor toekomstig experimenteel onderzoek (**Hoofdstuk 6**). Het proefschrift eindigt ten slotte met een verhandeling over het toekomstige gebruik van onze resultaten en modellen, en geeft een toekomstvisie voor onderzoek in de systeembiologie van antibiotica-producerende microben (**Hoofdstuk 7**).

Urdu summary

اِسٹرپٹومَأسِس پرجَاتیوں کو اِک بڑِی تعدَاد میں اَہم معَالجَاتی مرکبَات پیدَا کرنے کِی صلَاحیَت کی وجہ سے اَکثر طور پر «اِنٹیبَایوٹِک کَارخَانہ» کہَا جَاتَا ہے. وُہ اِکٹِنومَاءسِٹَالِس ترتیب سے تعلق رکھتے ہیں، جو کے حیَاتیَاتی طور پر جِنوم سَاءز میں اِختلَافَات دِکھلَانے کے لئے، پیتھوجنوسِٹی، مَاحولیَاتی طَاق میں اِختلَافَاتِ نظر، اَور کچھ ذَات کی مختلِف سِکنڈری مِٹَابولَاءٹس پیدَا کرنے کِی صلَاحیت سے بہُت متنوع ہے. یہ مقَالہ جینَس اِسٹرپٹومَأسِس اَور اُسکے رشتیدَاروں کو متعَارُف کرَانے کے سَاتھ مَاڈلِنگ ٹِکنولوجی کَا اِستِمعمَال اُن کی اِستحَالِی اَفعَال کو بِیَان کرنے کی تجزیہ سے شروع ہوتَا ہے (بَاب ۱). ہمنے اِنٹیبَایوٹِک پیدَا کرنے وَالے دو مَاڈل جرَاثیم اِسٹرپٹومَأسِس سِیلِیکُلُر (بَاب ۲) اَور اِسٹرپٹومَاسِس کلیوُلِگرَس (بَاب ۳) کِی اِستحَالِی نِظَام کَا مُطَالعہ کیَا، اَور کَمپیوٹر کے زریَہ اُن کِی اِنٹیبَایوٹِک پیدَا کَرنے کے نِظَام کِی تحقِیقَات کِی. یہ سمجھنیکیلئے کے کِس ترھ سے اِنٹیبَایوٹِک پیدَا کرنے وَالِی پرجَاتیَاں اِکٹِنومَاءسِٹَالِس گروپ کِی دوسری پرجَاتیوں سے نَسلی نوعی طور پر متعلِق ہیں، ہم نے اَیک جَامع نَسلی نوعی درخت کِی تعمِیر کِی، اَور جنوم سِکوِنس سے مُکمَّل طور پر نَسلی نوعی درخت کو حل کرنیکیلئے اَیک مَظبوت عَام اَستِعمَال کِی قَابِل نُقطہ نظر کِی بھِی تعمِیر کِی (بَاب ۴). نَسلی نوعی مطَالعہ کے نتَاءیج نے بڑے پیمَانے پر مِٹَابولِک موڈَل کِی بُنیَاد کو قَایم کِیَا، اَور ہم نے اُس گروپ کے اَرکَان کے درمیَان مِٹَابولِک اَور ٹوپولوجِکَل اَختِلَافَات اَور مشترکَات کِی شنَاخت بھِی کِی (بَاب ۵). اُس کے اِلَاوہ، نَسلی نوعی معلومَات کے سَاۃ جین اِکسپریشن کو جوڑ کر کے ہم نے اِسٹرپٹومَاءسِس

سِیلِیکُلر کے یتِیم جینس کو مستَقبِل میں مزید تجُربَاتِی مُطَالعہ کِی لئے ترجیحِی درجہ دیَا (بَاب ۶). آخِر میں مقَالہ کَا اِختِتَام ھمَارے نتَاءیج اَور مستَقبِل میں موڈَلس کے اَستِعمَال پر بَات چیت، اَور اُسکے اِلَاوہ اِنٹیبَایوٹِک پیدَا کرنیوَالے جرَاثِیموں کِی سِسٹمس بَایولوجی کے زریعے مزید تَحقِیق کے لئے کچھ نُقطہ نظَر کے ترتِیب کے سَاة ھوتَا ھے.

Acknowledgements

It gives me great pleasure in expressing my gratitude to all those people who have supported me and had their contributions in making this thesis possible. First and foremost, I must acknowledge and thank The Almighty Allah for blessing, protecting and guiding me throughout this period. I could never have accomplished this without the faith I have in the Almighty.

I express my profound sense of reverence to my supervisor and promoter Prof. Dr. Rainer Breitling, for his constant guidance, support, motivation and untiring help during the course of my PhD. His in-depth knowledge on a broad spectrum of different Bioinformatics and Systems biology topics has been extremely beneficial for me. He has given me enough freedom during my research, and he has always been nice to me. I will always remember his calm and relaxed nature, and the way he asks "YES! How can I help you?", whenever I enter his office. I am thankful to the Almighty for giving me a mentor like him.

I express my deepest gratitude to my second promoter and the Head of the Groningen Bioinformatics Center Prof. Dr. Ritsert C. Jansen for boosting my morale throughout the course of research. He has always been caring, a source of wisdom and motivation. He is a great leader.

I would like to thank Prof. Dr. Eriko Takano for helping me in Microbiology, and being available to guide me in my all projects and publications. Being a Mathematical modeler, I always find her comments, questions and suggestions in the manuscripts very challenging and I always felt very relaxed after answering her concerns.

I thank Elena and Marnix for being very good friends and brilliant collaborators. Their critical remarks and suggestions have always been very helpful in improving my skills and for strengthening our manuscripts. Both of them have agreed to become my paranymphs, and that certainly makes me feel confident during my defence. A special thank to Marnix for translating the thesis summary in to Dutch.

One person who has always been ready to help me was our secretary Klazien Offens. She took care of all non-scientific works, including official procedure of PhD promotie. Thank you Klazien for all your support!

It is my pleasure to acknowledge all my current and previous colleagues in GBiC, specially Marnix, Elena, Yang, Bruno, Andris, Richard, Frank, Anna, Lying, Nino, Rene, Joeri, Danny, Morris, Martijn, Jingyuan, Rudi, Francien and Gonzalo for their enormous support and providing a good atmosphere in the lab. I will always be grateful to them for helping me to develop the scientific approach and attitude. I would like to thank my colleagues in the University of Glasgow, Fiona and Andris for helping me in latex!

I also would like to acknowledge my current supervisors Dr. W. koopman and Dr. P. Willems, and my colleagues at the CSBB, UMC Nijmegen.

It's very important to have a nice social environment out side the lab. The city, Groningen, itself has taught me several good lessons and also has given me several good friends. In Groningen, very rarely I have felt that I am staying away from my home. In this regard, I would like to thank senior Indian PhD students of RuG including Ratna, Deepa, Ranjeet, Shireesha, Raj, Sriram, Kodanda, Aneesh, Samta, Hans, Anil, Vinay, Fiaz and many others whose name is not mentioned here for creating a wonderful social platform in the name of GISA. Inspired by the enthusiasm of GISA members, I started participating in GISA activities, and also served as a cultural secretary of GISA for the year 2009.

During my stay in Groningen, weekends were usually the times I recharged myself by interacting with the friends around. These interactions aided in improving my debating and culinary skills. I would like to thank my friends Fiaz, Sasanka, Jyothi, Yamini, Gopi, Harsh, Divya Raj, Divya Sasanka, Deepa, Bhushan, and Laksmi for the same. Although our debates typically ended in a inconclusive fashion and for which credit needs to be given to Fiaz for asking questions recursively. As a matter of fact, Jyothi answered all of them with same enthusiasm and full references. These interactions helped to improve my skills as a good listener, taught me how to organize my thoughts, and phrase my points.

I am grateful to my school and college teachers, Late. Maulvi Idrees, Master Rajendra babu, Prof. S. I. Ahson and Prof. Ram Ramaswamy, who laid seeds of enthusiasm and passion in my pursuit of knowledge.

I can't imagine my current position without the love and support from my family. I thank my parents, Mr. Md. Zaheer Alam and Mrs. Sajdah Khatoon, for striving hard to provide a good education for me and my siblings. I always fall short of words and felt impossible to describe their support in words. If I have to mention one thing about them, among many, then I would proudly mention that my parents are very simple and they taught me how to lead a simple life. I would simply say, "Amma, Abba you are great!".

My sisters Nazli aapa, Rubi aapa and Sazli, and brothers, Tanweer Alam, Tauseef Alam and Tafseer Alam, are great friends of mine. I would like to thank them for guiding me throughout. My sisters-in-law Ghazala and Tahseen, and brothers-in-law Tara bhai, Firoz bhai and Ijtaba Azam were also supportive. I acknowledge Zainab for her efforts in proof reading Urdu translation of the summary of the thesis.

Thanks to Hassan bhai for taking care of our family in the village. My cousins Affan Badar, Sufian Badar and Akhtar Nehal have always encouraged me to perform better. From my childhood, my uncles Late. Dr. Zafar Alam sb., Late. Badre Alam sb., Late. Shamshuzzuha sb., Mahboob akhtar sb., and Prof. Umar Ahsan sb. were very encouraging. My parents, brothers, sisters and uncles are always excited to hear my success and that inspires me to perform better and be successful. I acknowledge my entire family for providing us a very educated atmosphere in our village.

Last but not the least I would like to remember and thank my paternal grandfather Late. Hafiz Abdur Razzaque sahab and maternal grandfather Late. Maulvi Mohibbul Hasan sahab for their prayers. I strongly believe that their prayers played very important role in my life. They are not here with me anymore but their prayers are, and Insha-Allah Taa'la everything will be good (Ameen).

Publications

- **Alam MT**, Medema MH, Takano E, Breitling R: *Comparative genome-scale metabolic modeling of actinomycetes: the topology of essential core metabolism.* **(submitted)**

- **Alam MT**, Takano E, Breitling R: *Prioritizing orphan proteins for further study using phylogenomic and gene expression profiles in* Streptomyces coelicolor. **(submitted)**

- Medema MH*, **Alam MT***, Breitling R, Takano E: *The future of industrial antibiotics production: from random mutagenesis to synthetic biology.* **Bioengineered Bugs 2011 (in press)**
  *Equal contribution

- Medema MH*, **Alam MT***, Heijne W, Berg MVD, Müller U, Trefzer A, Bovenberg RAL, Breitling R, Takano E: *Genome-wide gene expression changes in an industrial clavulanic acid overproduction strain of* Streptomyces clavuligerus. **Microb. Biotechnol. 2011, 4(2):300–5.**
  *Equal contribution

- **Alam MT**, Merlo ME, The STREAM consortium, Hodgson DA, Wellington EMH, Takano E, Breitling R: *Metabolic modeling and analysis of the metabolic switch in* Streptomyces coelicolor. **BMC Genomics 2010, 11:202**

- **Alam MT**, Merlo ME, Takano E, Breitling R: *Genome-based phylogenetic analysis of* Streptomyces *and its relatives.* **Mol. Phylogenet. Evol. 2010, 54:763–772**
  [Faculty of 1000 Biology **"Must read"** paper]

- Medema MH, Trefzer A, Kovalchuk A, Berg MVD, Müller U, Heijne W, Wu L, **Alam MT**, Ronning CM, Nierman WC, Bovenberg RAL, Breitling R, Takano E: *The sequence of a 1.8-Mb bacterial linear plasmid reveals a rich evolutionary reservoir of secondary metabolic pathways.* **Genome Biol. Evol. 2010, 2:212-224**
  [Faculty of 1000 Biology **"Must read"** paper]

- Nieselt K, Battke F, Herbig A, Bruheim P, Wentzel A, Jakobsen OM, Sletta H, **Alam MT** et al.: *The dynamic architecture of the metabolic switch in* Streptomyces coelicolor. **BMC Genomics 2010, 11:10**
  [BMC **Highly accessed** paper]

Curriculum vitae

Mohammad Tauqeer Alam was born on 9 January 1981 in the Village Nazra in the District Madhubani, India. In 1995 he finished secondary school at his village school (Bihar School Examination Board, Patna) and went to Karim City College Jamshedpur to study science. He obtained a B.Sc. in computer science from Magadh University, and in 2003 he qualified for the M.Sc. Bioinformatics at Jamia Millia Islamia University, New Delhi. He secured the top rank in his M.Sc. class and received a gold medal from the University. In 2005, he continued to study Bioinformatics at Jawaharlal Nehru University, New Delhi, India. In January 2007 he joined the Groningen Bioinformatics Center, University of Groningen, The Netherlands as a Ph.D. student under the supervision of Prof. Dr. Rainer Breitling and Prof. Dr. Ritsert C. Jansen. There, he studied metabolism of actinomycetes species by constructing *in silico* models. In the final year of his Ph.D., in March 2010, he moved to the University of Glasgow, where he finished his project under the supervision of Prof. Breitling.

During his Ph.D research, two of his papers were evaluated by the Faculty of 1000 Biology as *"Must read"* paper, and one paper was permanently marked as *Highly accessed* by BioMed Central.