GENOMIC ANALYSIS OF REGULATORY MECHANISMS INVOLVED IN SECONDARY METABOLITE PRODUCTION

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Marlene Castro-Melchor

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Wei-Shou Hu, Adviser

September, 2010

Acknowledgements

The successful culmination of my Ph.D. studies would not have been possible without the tremendous help and unconditional support of my mother, MMM. She endured not only Minnesota winters, but me. Although from a distance, my sister Liliana always represented a bright spot during this time.

My research projects, which gave me the opportunity of interacting with so many people, are the result of the visionary guidance of my adviser, Prof. Wei-Shou Hu. As everybody that knows him, I am impressed by his energy and enthusiasm. My research projects put me in contact with wonderful collaborators such as Prof. David Sherman, Prof. George Karypis, Prof. Rachel Chen, and Dr. Anne Ruffing. Dr. Zheng Jin Tu, from the Minnesota Supercomputer Institute was an invaluable source in my bioinformatics' endeavors.

The members of the Hu group are the ones that made my experience in graduate school what it was. As part of the Streptomics group, I mostly worked with Dr. Karthik Jayapal and Dr. Salim Charaniya, but helpful hands extended from members before them, such as Dr. Wei Lian. I am thankful to past members Dr. C. M. Cameron, Dr. Patrick Hossler, Dr. Fernando Ulloa-Montoya, Dr. Joon Chong Yee, and Dr. Anne Kantardjieff for sharing their experiences and providing me with understanding. I am most thankful to Siguang Sui, who is the best lab and office mate ever! He was always there to help me take samples and when my computer crashed. I would like to acknowledge the help of Lee Fredrickson and Chia-Chun Lee, who prepared their share of all sorts of media.

Thanks to Lucy Camarena and Dorian Haro, who keep my sense of connection with Guadalajara. From my hometown, I know I will always have the true friendship of Edgar, David, and Mayra, and the mentorship of my former boss Dr. Rogelio Lopez-Herrera. Thanks to Maribel Nuñez and Dr. Alma Rodriguez, who made me laugh like no one else!

Although we couldn't see each other that often, I am deeply thankful to Stephan Cameron, who is the person that listened to my most personal frustrations and disappointments. I would like to thank Huong Le, without whom I wouldn't have made it through. Together we not only picked blueberries, strawberries, and apples, but also shattered dreams.

Dedication

To my mother, MMM, and the memory of my father.

Abstract

Many secondary metabolites have beneficial uses for humans. In addition to their use as antibacterial and antifungal agents, secondary metabolites have been used as immunosuppressants, anti-tumor agents, and antiparasitics. Most of the secondary metabolites known today are produced by filamentous fungi or by members of the *Streptomyces* genus. Production of secondary metabolites by microorganisms involves a complex, dynamic system, with interconnected elements acting at different levels.

Diverse tools were used in this work to explore regulation of secondary metabolite production, mostly in *Streptomyces*. The tools have a common characteristic: they either generate large amounts of data, or require large amounts of data.

Regulation of secondary metabolite production in *Streptomyces coelicolor* was analyzed at the genome level, by using network modules inferred from a large transcriptome dataset. The upstream sequence of the elements in the network modules was searched for the presence of consensus sequences, and these results combined with information on known interactions, binding sites, and functional relatedness. The combination of this information resulted in a set of twenty networks that have a high likelihood of representing true interactions and represent a starting point for further experimental studies.

The characteristics of high productivity were analyzed by comparing the genomes of two strains of the clavulanic acid producer *Streptomyces clavuligerus*. One of the strains is a high producer of clavulanic acid. Next generation sequence data was used to perform a genome-wide screening to identify all the differences between the two genomes. In addition to mutations in genes involved in β-lactam antibiotic production or their upstream region, structural differences were detected between the two strains.

Next generation sequencing technologies were also used to assemble a draft genome for the curdlan producer *Agrobacterium* sp. ATCC 31749. Curdlan production mimics that of secondary metabolites, it is triggered under starvation conditions.

These varied approaches exemplify some of the paths that can lead to a better understanding of secondary metabolism and its regulation.

Table of Contents

List of Tables

List of Figures

Chapter 1
Introduction

1.1 The relevance of secondary metabolites

Humans have harnessed the potential of microorganisms before we even knew they existed. In Babylonia, beer was produced by the action of yeast converting sugar to alcohol, Egyptians ate leavened bread thanks to the CO_2 generated by the action of brewer's yeast, and wine was being made in Assyria by 3500 BCE (Demain and Fang 2000). But it was not until the seventeenth century that Antonie van Leeuwenhoek first described microorganisms. The transforming activity of bacteria and yeast was first made clear in food products, thanks to work by Pasteur, among others. The golden era of antibiotics began in 1929, when the production of penicillin by mold was discovered.

The process to establish the large scale production of penicillin initiated a series of collaborations between pharmaceutical companies, academic laboratories, and government agencies, and was at the center of what would later become drug discovery and research programs. Process improvement, strain selection and screening were all steps leading to the successful production of penicillin. Genetics contributed the proof that mutations were responsible for some of the increases in penicillin titers. Later on, chemistry contributed semi-synthetic production, by modifying prototypic structures to obtain new compounds or compounds with improved or novel effects (Drews 2000).

Streptomycin, the first effective antibiotic against tuberculosis was discovered in 1944 (Raju 1999). It was the first successful antibiotic produced by an actinomycete. Actinomycetes produce approximately two-thirds of natural antibiotics (Okami and Hotta 1988), and members of the genus *Streptomyces* are the producers of 75% of commercially and medically useful antibiotics (Miyadoh 1993).

Most natural antibiotics are produced as secondary metabolites, that is, they have no effect on growth. In nature, however, they contribute to the survival of the producing strain. After the success of penicillin and streptomycin, secondary metabolites were mostly screened for their antibacterial and antifungal properties, but in the 1960's, researchers in Japan screened secondary metabolites for activities other than bacterial and fungal inhibition. These broader screens resulted in the identification of extremely useful secondary metabolites with anti-enzyme activity. In addition to enzymes, receptors were suggested as targets for drugs by yet another discipline, biochemistry. But secondary metabolites have functions outside their antimicrobial properties. They are used as differentiation effectors, sexual hormones, and agents of symbiosis (Demain and Fang 2000).

Recombinant DNA technology has contributed to improving secondary metabolite production by modified organisms. Overproduction of secondary metabolites has been achieved by the amplification of genes coding for enzymes intervening in biosynthetic pathways, and non producing organisms have been made producers by the introduction of all the genes responsible for the production of specific secondary metabolites. This last process has been facilitated by the realization that genes involved in secondary metabolite biosynthetic pathways are generally clustered in the genome. In addition, unnatural natural products have been obtained by introducing antibiotic synthetases into producers of other antibiotics. Polyketides are particularly amenable to combinatorial chemistry, as their biosynthesis involves modules responsible for full cycles of elongation or group modification (Cane, Walsh et al. 1998).

Even though huge progress has been made in characterizing and understanding secondary metabolites in the last 80 years, many questions remain unanswered. The biosynthesis of secondary metabolites is complex and involves regulation at different levels. After a period of decline in the interest of the pharmaceutical industry for the discovery of natural products, the unrealized expectations from high-throughput screening of synthetic chemical libraries has prompted a renewed interest for natural products (Koehn and Carter 2005).

After the contributions made by microbiology, biology, chemistry, biochemistry, and molecular biology to the understanding of secondary metabolite production, genomics is starting to make its impact too. The availability of the sequence of the full genome of secondary metabolite producers has started a new era in the study of secondary metabolism. Inexpensive next generation sequencing is providing us with more and more genomes of secondary metabolite producers, and creating the possibility of even sequencing different strains. Whole genome differences can now be determined for two organisms, and post-genomic fields like systems biology could help in reconstructing the metabolism of most species (Li and Vederas 2009). Synthetic biology's goal is to design and construct new biological parts, devices and systems (Porcar 2010), and genes involved in secondary metabolite production could provide some of the most valuable parts for the design of complex molecules.

1.2 Scope of this thesis

The focus of this work is in the regulation of secondary metabolite production in bacteria. The work used different tools for the study of different organisms, reflecting the maturation level that the study of such organisms has reached. Two streptomycetes were used in this study, the model actinomycete *Streptomyces coelicolor* and the industrially relevant *Streptomyces clavuligerus*. In addition, work was done on *Agrobacterium* sp. ATCC 31749, producer of an exopolysaccharide.

Streptomyces coelicolor produces several antibiotics, two of which are colored. The availability of its complete genome sequence in 2002 contributed to an increase in the number of studies focusing in understanding the regulation of secondary metabolite production in this organism. In our lab, Dr. Lian constructed a whole-genome cDNA microarray for this organism and, together with Dr. Jayapal, characterized the transcriptome of several mutants constructed by Dr. Kyung. Dr. Mehra focused on modeling a particularly interesting system in this organism that involves a bacterial hormone and has an impact on antibiotic production.

As transcriptome data for this organism accumulated, Dr. Charaniya realized the opportunity to use it for the prediction of operons. Later on, he used the transcriptome data for elucidating the genome-wide regulatory network of this organism. This work focuses on combining information from those predicted networks with other bioinformatics' approaches based on genomic features to add biological interpretation to them. By assessing these networks with known interactions, clusters, and motifs, we were able to appraise the value of systems approaches in providing additional target candidates for known regulators and suggesting targets for as yet unstudied ones.

Contrary to the case of *Streptomyces coelicolor*, the genomic information available for *Agrobacterium* sp. ATCC 31749 was almost none existing. The group of Dr. Chen at the Georgia Institute of Technology was working on engineering sugar polymer synthesis in *Agrobacterium* sp. ATCC 31749. Their work, however, was limited to genomic information of the plant pathogen *Agrobacterium tumefaciens*, which does not produce the exopolysaccharide of their interest. In this collaborative work, we used next generation sequencing to construct a draft genome for *Agrobacterium* sp. ATCC 31749. The genome was compared to that of other *Agrobacterium* and *Rhizobium* species, and open reading frames were predicted and annotated. The information generated by this work was later used to design an oligonucleotide microarray which is being used to further study exopolysaccharide production in *Agrobacterium* sp. ATCC 31749.

Next generation sequencing was also used to study the characteristics of high productivity, in this case by comparing two strains of the clavulanic acid producer *Streptomyces clavuligerus*. Our group had previously worked with *Streptomyces clavuligerus*, but focused on *Streptomyces coelicolor* when the genome of the later became available. Work by Dr. Kyung studied the effect of the regulator *ccaR* on the production of cephamycin C, and later on Dr. Lian characterized the effect of the two-component system *cbsA* in the production of both cephamycin C and clavulanic acid. The tools have changed since then though, and in this work rather than studying a single gene, a whole-genome mutation screening revealed structural differences and critical point mutations between two strains.

1.3 Thesis organization

This thesis is organized in seven chapters plus three appendixes. Chapter 2 presents an overview of the genus *Streptomyces*, with an emphasis on the production of secondary metabolites and their regulation. Chapter 3 reviews next generation sequencing (NGS) technologies and the tools used in this work for handling NGS data. Chapter 4 describes the work done on reverse engineering the regulatory network of *Streptomyces coelicolor* and the different layers of information that were used to assess them. Network modules with high probability of representing true interactions are presented in detail. Chapter 5 summarizes the work done to assemble a draft genome for *Agrobacterium* sp. ATCC 31749, its comparison to close organisms, and the subsequent bioinformatics' work performed to predict and annotate protein-coding genes. Chapter 6 describes the comparison of two strains of *Streptomyces clavuligerus* using next generation sequencing. The results of genome-wide mutation screening to single out those most likely involved in the high antibiotic production phenotype are presented. Chapter 7 summarizes the results and conclusions of this work.

These chapters are followed by three appendixes. Appendix A and B provide supplementary information to chapters 4 and 6, respectively. Appendix C presents a small project exploring the bioprocess characteristics contributing to high productivity and product quality in the production of a monoclonal antibody by mammalian cell culture. This project used data mining techniques on 51 production runs and data on more than 20 process parameters.

Chapter 2
Background

2.1 Summary

This chapter is an overview of the genus *Streptomyces*. The production of secondary metabolites by this genus is stressed, including its complex regulation. Two members of this genus are reviewed in more detail: the model organism *Streptomyces coelicolor* and the industrially relevant *Streptomyces clavuligerus*.

2.2 The genus *Streptomyces*

Gram-positive bacteria include two main branches: the low G+C organisms, and the high G+C organisms. The high G+C organisms are the actinomycetes (Kieser, Bibb et al. 2000). Actinomycetes live in soil, where it is estimated that one gram of rich soil can contain millions of colony-forming units of these organisms (Donadio, Sosio et al. 2002). Among actinomycetes, the G+C content ranges from 54% in Corynebacteria to more than 70% in *Streptomyces* (Chater 2006). *Streptomyces* undergo a complex life cycle, including filamentous vegetative growth, aerial hyphae formation, and differentiation of the aerial hyphae into spores (Paradkar, Trefzer et al. 2003). Spores assist the spread and persistence of *Streptomyces* in soil (Kieser, Bibb et al. 2000). *Streptomyces* are renowned for their secondary metabolite production. Some compounds produced by *Streptomyces* have antibacterial and/or antitumor activity, some examples are given in Table 2.1.

Table 2.1. Some secondary metabolites produced by *Streptomyces* with antibiotic or antitumor activity (Kieser, Bibb et al. 2000; Martin, Casqueiro et al. 2005).

Class	Examples
β-lactams	Cephamycin C Clavulanic acid Nocardicin
Anthracyclines	Daunorubicin Doxorubicin Tetracenomycin
Aureolic acid group	Chromomycin Muthramycin Olivomycin
Glycopeptides	Bleomycin Pleomycin
Tetracyclines	Chlortetracycline Oxytetracycline

Streptomyces are also characterized by long linear chromosomes and in several cases long plasmids have also been observed. Plasmids do not carry essential genes and their replication is independent of the chromosome and uncoupled from the cell cycle (Egan, Fogel et al. 2005). Linear plasmids contain a centrally located origin of replication. Replication of complete chromosomes and plasmids is achieved by the action of proteins bound to the 5' end of the double stranded telomeres. *Streptomyces*' linear chromosomes and plasmids contain conserved palindromic sequences at their edges, known as terminal inverted repeats.

Streptomyces' chromosomes are genetically unstable. Since no essential genes are located at the end of chromosomes, *Streptomyces* are able to tolerate large deletions in that region without any effect in viability. Deletions can affect about 0.5% of germinating spores (Hopwood 2006). Large deletions of up to two million base pairs (Mb) have been identified at or near the ends of chromosomes (Chen, Huang et al. 2002).

2.3 Secondary metabolite production in *Streptomyces* and its regulation

Secondary metabolites are nonessential for growth, and are usually synthesized at the end of the exponential growth phase. Although secondary metabolite production is not essential to the producing organism, it has been suggested that it confers a competitive advantage. In the case of *Streptomyces* their natural environment is soil, one of the most competitive and changing environments. Genes for the production of secondary metabolites in *Streptomyces* are found in clusters, which include not only the biosynthetic genes, but also those for resistance and export of the secondary metabolites. The expression of efflux genes is coordinated with that of biosynthetic genes. In *Streptomyces*, resistance to antibiotics is conferred by ABC transporters and major facilitator superfamily (MFS) transporters (Martin, Casqueiro et al. 2005).)

2.3.1 *Streptomyces* antibiotic regulatory proteins (SARPs)

Antibiotic production control in *Streptomyces* involves multiple layers, including gene products involved in morphological differentiation, pleiotropic regulation, and pathway specific regulation. Most pathway specific regulators in *Streptomyces* belong to the *Streptomyces* antibiotic regulatory proteins (SARPs) family (Wietzorrek and Bibb 1997). These regulators bind to direct heptameric repeats that sometimes overlap the -35 region. Examples of SARPs include ActII-ORF4 and RedD in *Streptomyces coelicolor*, DnrI in *Streptomyces peucetius*, and CcaR in *Streptomyces clavuligerus*. These proteins are the pathway specific regulators of the antibiotics actinorhodin, undecylprodigiosin, daunorubicin, and the β-lactams cephamycin C and clavulanic acid, respectively.

2.3.2 γ-butyrolactones

Hormone-like signaling molecules known as γ-butyrolactones can play a role in morphological differentiation and are associated with onset of antibiotic production in *Streptomyces* (Miguelez, Hardisson et al. 2000). This type of molecules has been reported in *Streptomyces griseus*, *Streptomyces viridochromogenes*, *Streptomyces bikiniensis*, *Streptomyces cyaneofuscatus*, *Streptomyces virginiae*, *Streptomyces* sp. FRI-5, and *Streptomyces coelicolor*. Three types of γ-butyrolactones have been defined based on their side chain: VB-type, IM-2 type, and A-factor type. VB-A to VB-E are produced by *Streptomyces virginiae* and contain a 6-α-hydroxy group, IM-2 is produced by *Streptomyces sp* FRI-5 and contains a 6-β-hydroxy group, and A-factor is produced by *Streptomyces griseus* and contains a 6-keto group. In *Streptomyces coelicolor* the γ-butyrolactone SCB1 was found to have an effect on the production of the two pigmented antibiotics actinorhodin and undecylprodigiosin. Although a gene encoding a γ-butyrolactone receptor has been described in *Streptomyces clavuligerus* (Kim, Lee et al. 2004), the corresponding butyrolactone remains unknown.

2.3.3 Two-component systems (TCSs)

Two-component systems (TCSs) are a signal transduction mechanism in which a histidine kinase undergoes autophosphorylation in response to a detected signal and then catalyzes phosphotransfer to a response regulator. The phosphorylated response regulator usually binds DNA and activates transcription of specific genes.

In most cases, genes encoding a pair of histidine kinase and response regulator are next to each other in the genome and are expressed as a single transcript. Autoregulation, that is, the binding of the phosphorylated response regulator to the promoter region of the TCS operon has been detected in some cases.

2.3.4 The stringent response

In response to extreme nutrient limitation, bacteria undergo a stringent response consisting of the accumulation of guanosine tetraphosphate, ppGpp, and a reduction in GTP levels. ppGpp binds to RNA polymerase, which results in a global reduction of transcription, and induction of stress response genes. In *Streptomyces coelicolor*, ppGpp plays a role in triggering the onset of antibiotic production, as suggested by the correlation between ppGpp accumulation and transcription of the pathway-specific regulators *redD* and *actII-ORF4* (Ochi 2007). In *Streptomyces clavuligerus*, however, there is a lack of correlation between ppGpp levels and the expression of pathway specific regulators (Liras, Gomez-Escribano et al. 2008).

2.4 Streptomyces coelicolor

2.4.1 The *Streptomyces coelicolor* genome

As the model organism for actinomycetes, *Streptomyces coelicolor* was the first organism in this group with a sequenced genome. In 2002, when the *Streptomyces coelicolor* genome was sequenced by (Bentley, Chater et al. 2002), it represented the largest bacterial genome sequenced thus far (8.7 Mb). A total of 7825 coding sequences were predicted, with 5% of them corresponding to genes coding for secondary metabolites, arranged in 23 clusters, and 10% of the predicted coding sequences corresponding to regulators. The regulators include 63 sigma factors, 84 histidine kinases, and 80 response regulators. Analysis of the chromosome revealed a central core containing housekeeping genes, and two arms with a high fraction of secondary metabolite clusters.

2.4.2 Morphological differentiation in *Streptomyces coelicolor*

Mature colonies (Figure 2.1) of *Streptomyces coelicolo*r on agar are multicellular, containing vegetative mycelia which grow on and into the agar, aerial mycelia which produce spores, and spores which are resistant to desiccation (Hopwood, Chater et al. 1973). Two types of mutants have contributed greatly to the understanding of morphological differentiation in *Streptomyces coelicolor*. The first type is known as *bld* for "bald", the phenotype characterized by the lack of aerial hyphae formation (Figure 2.2). The *bld* genes regulate initiation of growth of aerial hyphae in a complex signaling cascade transmitted by the sequence *bldJ* → *bldK*/*L* → *bldAH* → *bldG* → *bldC* → *bldD*/*M* → *ram*. The second type of mutants is known as *whi* for "white", the phenotype characterized by the lack of mature gray spores. Among the genes in this cascade are *whiA*, *whiB*, *whiD*, *whiE*, *whiG*. *whiH*, and *whiI*. These two cascades converge with the repression of *whiG* by BldD (Chater and Chandra 2006). A third pathway is involved in morphological differentiation in *Streptomyces coelicolor*. The sky pathway regulates the expression of chaplins and rodlins, which help in the formation of a rodlet layer that provides hydrophobicity to aerial hyphae and spores. While some *bld* mutants exhibit a conditional phenotype, chaplin deletions result in unconditional mutations which fail to grow regardless of media composition (Claessen, Stokroos et al. 2004; Claessen, de Jong et al. 2006).

Secondary metabolite production is linked to morphological differentiation, occurring at the same time as aerial hyphae are formed. Links to primary metabolism have also been suggested, as some *bld* mutants are affected in carbon catabolite repression (Chater 2001).

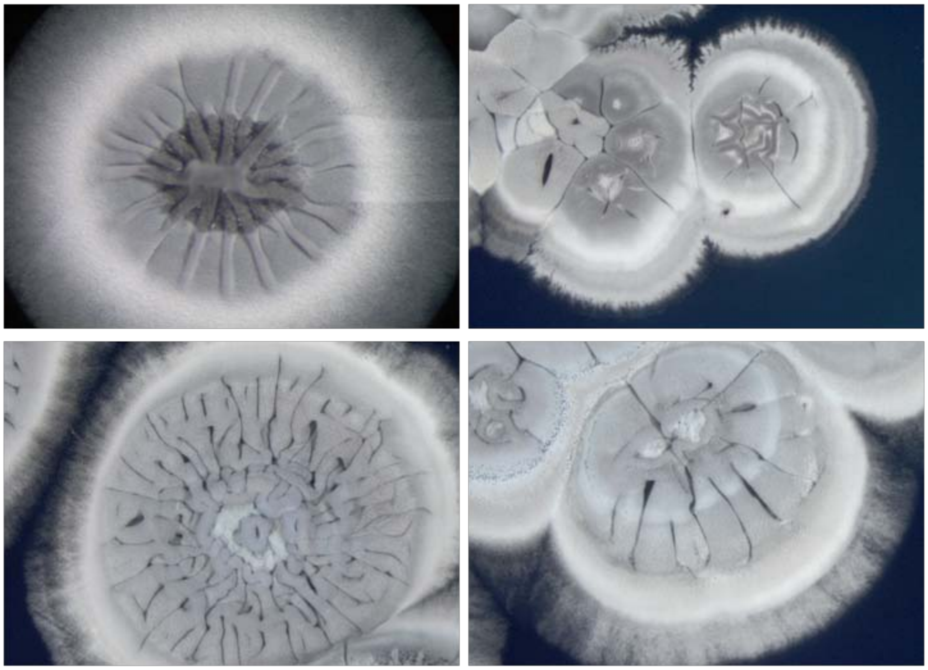


Figure 2.1. Mature *Streptomyces coelicolor* colonies. Aerial hyphae are visible at colony edges.

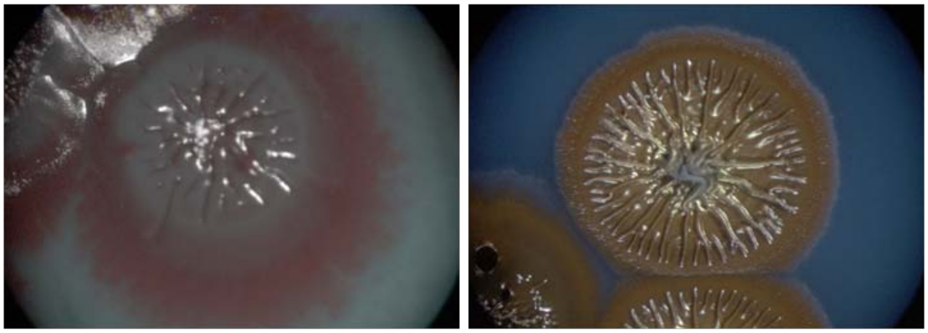


Figure 2.2. *Streptomyces coelicolor* mutants exhibiting "bald" phenotype. The blue pigment in the background is the antibiotic actinorhodin.

2.4.3 Secondary metabolites produced by *Streptomyces coelicolor*

Streptomyces coelicolor produces two pigmented antibiotics: actinorhodin, known simply as ACT, and undecylprodigiosin, also known as RED. Other compounds produced by *Streptomyces coelicolor* that have been studied include the calcium-dependent antibiotic (CDA), methylenomycin (Mmy), and the gray spore pigment (WhiE). Methylenomycin is produced by genes located on the giant linear plasmid SCP1. The strain used in this project, *Streptomyces coelicolo*r M145, is a plasmidless strain and thus does not produce methylenomycin.

Actinorhodin is a polyketide derived from a type II polyketide synthase (PKS) that appears red under acidic conditions and blue under basic conditions. ACT is excreted into the medium and into aqueous droplets (Figure 2.3) on the hydrophobic surface (Thompson, Fink et al. 2002). The ACT cluster extends for 21 Kb, from locus SCO5071 to SCO5092. The actinorhodin cluster activator protein ActII-ORF4 belongs to the SARP family. ActII-ORF4 binds to the intergenic region of the divergently transcribed *actVI-ORF1* and *actVI-ORFA* which are late-step genes in the ACT production pathway, and to the intergenic region of the early-step ACT cluster genes *actIII* and *actI*. The binding to these two regions is not of the same strength, as higher affinity was detected for the intergenic region of the late genes (*actVI*) (Arias, Fernandez-Moreno et al. 1999). In addition to ACT, the gray spore pigment WhiE is encoded by a type II PKS system (Moore and Piel 2000).

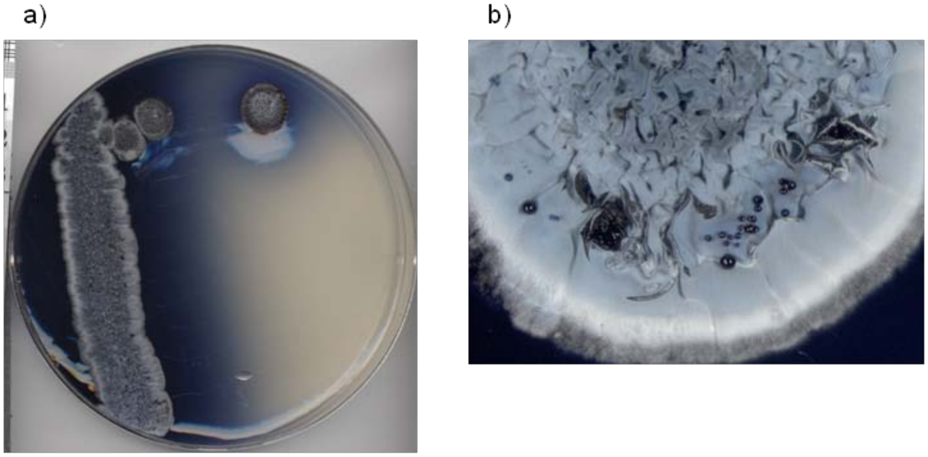


Figure 2.3. a) *Streptomyces coelicolor* colonies in a Petri dish. The diffusion of the blue colored antibiotic ACT is clearly seen. b) Colony of *Streptomyces coelicolor* showing aqueous droplets of ACT secreted into the hydrophobic surface.

Undecylprodigiosin is the most abundant among the four prodiginines in the mixture produced by *Streptomyces coelicolor* (Tsao, Rudd et al. 1985). It is known as RED for its characteristic color (Figure 2.4). The RED biosynthetic cluster extends for 32 Kb, from locus SCO5877 to SCO5898. Early biosynthetic genes are centrally located and flanked by late step regulatory genes (Coco, Narva et al. 1991). Production of RED is complex, involving two pathway specific regulators encoded by *redD* and *redZ*. RedD belong to the SARP family. RedZ is an activator of *redD*. *redZ* contains a rare codon, UUA which is translated efficiently only by the transfer RNA (tRNA) encoded by *bldA*. *bldA* mutants present in addition to the bald phenotype, lack of production of RED, and lack of production of ACT. The pathway specific regulator *actII-ORF4* also contains a UUA codon (Takano, Tao et al. 2003).

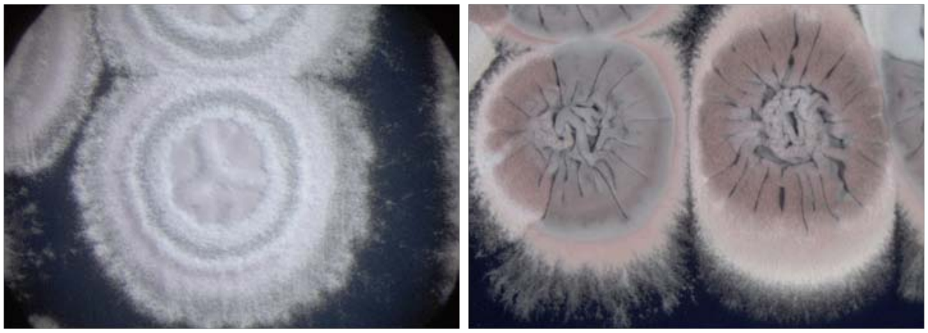


Figure 2.4. *Streptomyces coelicolor* colonies producing the red antibiotic undecylprodigiosing, which is noticeable as shades of pink in the colonies.

CDA is an acidic lipopeptide antibiotic which requires the presence of calcium ions, thus its name Calcium-Dependent Antibiotic (Lakey, Lea et al. 1983). The CDA cluster is one of the longest secondary metabolite clusters, both in terms of length and number of coding sequences. It extends for 83 Kb, from locus SCO3210 to SCO3249. The cluster is located within the core region of the chromosome, where housekeeping genes are concentrated (Hojati, Milne et al. 2002).

Recently, the deletion of *scbR2*, a homologue of the γ-butyrolactone receptor *scbR*, resulted in the detection of a compound with antibacterial activity, abCPK, and a yellow-pigmented secondary metabolite, yCPK (Gottelt, Kol et al.). *scbR2* is within the *cpk* gene cluster, which was until now an orphan type I polyketide synthase gene cluster.

Polyketides are of interest in combinatorial chemistry for developing unnatural natural products. The goal of combinatorial chemistry is to increase structural diversity of natural products by following a strategy of mix-and-match, for which the more than 500 multicyclic aromatic polyketides that have been characterized in actinomycetes is particularly useful (Moore and Piel 2000).

2.4.4 Two-component systems in *Streptomyces coelicolor*

The *Streptomyces coelicolor* genome encodes 84 histidine kinases and 80 response regulators, forming 67 pairs of TCSs (Hutchings, Hoskisson et al. 2004). Some of the orphan response regulators include genes with a known effect in morphology and antibiotic production, e.g., *ramR*, *redZ*, *whiI*, and *bldM* (Willey, Willems et al. 2006). In some cases TCSs form part of signaling cascades, for example *ramR* is not transcribed in *bldA*, *bldB*, *bldH*, and *bldD* mutants, indicating its possible activation by the *bld* cascade.

Several TCSs have a pleiotropic effect on antibiotic production. For example, the TCS integrated by AbsA1 and AbsA2 regulates production of all four antibiotics (ACT, RED, CDA, and Mmy) in *Streptomyces coelicolor*. This TCS is located within the CDA biosynthetic gene cluster, although there is no evidence that it regulates transcription of the pathway specific regulator of this cluster, *cdaR* (Ryding, Anderson et al. 2002). AbsA does however affect the expression of the pathway specific regulators *actII-ORF4* and *redD* (Aceti and Champness 1998). Another TCS with pleiotropic effects is the one integrated by the histidine kinase AfsQ2 and the response regulator AfsQ1, which stimulate the production of ACT and RED (Ishizuka, Horinouchi et al. 1992).

Other TCSs with a more limited effect on antibiotic production have been reported, as is the case of the first TCS identified in *Streptomyces*, CutRS. Deletion of *cutRS* resulted in overproduction of ACT (Chang, Chen et al. 1996). Deletion of the TCS PhoRP also results in overproduction of ACT and RED in complex media (Sola-Landa, Rodriguez-Garcia et al. 2005).

2.4.5 Serine/Threonine (Ser/Thr) kinases in *Streptomyces coelicolor*

Ser/Thr kinases catalyze phosphorylation of target proteins in serine and threonine residues. Ser/Thr kinases are common in eukaryotes, but present in only some prokaryotes. Furthermore, the number of Ser/Thr kinases in a genome seems to correlate with complexity of life cycle. The genome of *Streptomyces coelicolor* encodes 34 Ser/Thr protein kinases, but few have been studied so far.

The best documented case is that of the Ser/Thr kinase AfsK. Activated AfsK catalyzes phosphorylation of the SARP AfsR, which then binds to the promoter region of *afsS* (Petrickova and Petricek 2003). AfsS is a pleiotropic regulator of antibiotic synthesis (Lian, Jayapal et al. 2008).

2.5 Streptomyces clavuligerus

2.5.1 The *Streptomyces clavuligerus* genome

Even though *Streptomyces clavuligerus* is of industrial importance, projects to sequence its genome weren't announced until 2008. By 2008 two genome projects appeared on the National Center for Biotechnology Information (NCBI) website, one by the Korea Research Institute of Bioscience and Biotechnology (KRIBB) and the other by The Broad Institute Genome Sequencing Platform. Although no preliminary data was available from KRIBB, a draft release was available at the Broad Institute by August 2008. This preliminary assembly consisted of 597 contigs and a total length of 6.7 million base pairs (Mb), which appeared small for a *Streptomyces* genome. By November 2009 a new draft assembly was available by the Broad Institute. The second release consisted of 1036 contigs and a total length of 9.1 Mb, which seems much more in the range of known *Streptomyces* genomes (8.5 – 10.1 Mb).

A third group at the University of Groningen, working in collaboration with the DSM Biotechnology Center of DSM Food Specialties, released a sequence for *Streptomyces clavuligerus* earlier this year (May 2010) (Medema, Trefzer et al. 2010). This release, which used Sanger sequencing and the aid of an optical restriction map for assembly, consists of 279 contigs and a total length of 8.7 Mb. The genome consists of a large linear chromosome of 6.7 Mb, which is small in *Streptomyces* terms, and a giant linear plasmid, pSCL4, of 1.8 Mb. The combined length of the chromosome and the plasmid is in the typical size range of *Streptomyces* genomes.

The DSM sequence includes 7281 coding sequences and an extremely high number of gene clusters involved in secondary metabolite production. The *Streptomyces clavuligerus* genome contains 23 secondary metabolite clusters in its chromosome and 25 more in the giant plasmid pSCL4.

2.5.2 β-lactam antibiotics

Numerous β-lactam antibiotics (penicillins and cephalosporins, including cephamycins) of medical importance are produced by mycelial bacteria and filamentous fungi. Among the bacterial producers, *Streptomyces clavuligerus* and *Nocardia lactamdurans* are used in the commercial production of β-lactam antibiotics. The fungi *Acrenomium chrysogenum* and *Penicillium chrysogenum* are also used in commercial scale preparation of these antibiotics. These antibiotics are synthesized from lysine, cystine, and valine (Rius and Demain 1997) and contain a β-lactam ring (Figure 2.5).

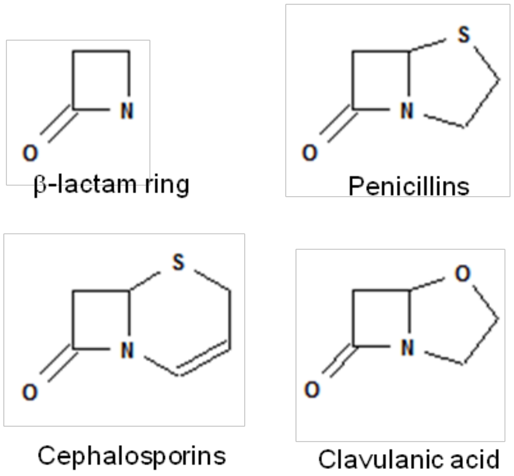


Figure 2.5. β-lactam ring and basic structure of β-lactam antibiotics, which contain this ring.

2.5.2.1 Cephamycin C

Cephamycins were the first group of β-lactam antibiotics discovered from actinomycetes (Liras and Demain 2009). Cephamycins are modified cephalosporins (Figure 2.6) and are active against penicillin-resistant bacteria (Liras 1999). Many of the steps in its biosynthetic pathway are common to those of penicillins and cephalosporins (Figure 2.7), including its synthesis from L-α-aminoadipic acid (α-AAA), L-cysteine, and L-valine. Cephamycin production by actinomycetes, however, requires additional genes to produce the precursor α-AAA, which is an intermediate in lysine formation in eukaryotes. Once α-AAA has been generated the next five steps in the biosynthetic pathway are similar to those for cephalosporin synthesis in *Acrenomium chrysogenum*. Intermediates include isopenicillin N, deacetoxycephalosporin C (DAOC), and deacetylcephalosporin C (DAC). The final steps, modifying DAC, are unique to actinomycetes.

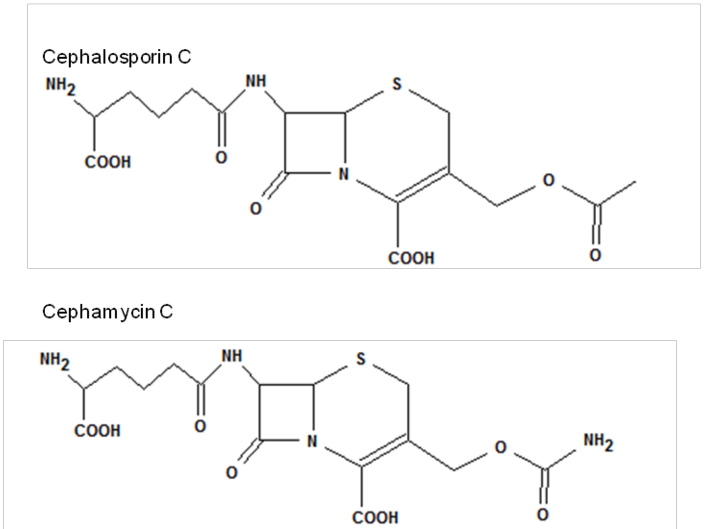


Figure 2.6. Structure of cephalosporin C and cephamycin C.

The cephamycin C cluster is next to the clavulanic acid cluster in *Streptomyces clavuligerus*, forming the β-lactam supercluster. The level of several enzymes in the pathway is influenced by the pathway-specific regulator CcaR, which also affects clavulanic acid production. ClaR, a regulator located on the clavulanic acid cluster also seems to affect cephamycin C production, although the mechanism has not been elucidated. The cephamycin C cluster is 36 Kb long, extending from locus SCLAV-4198 (*pcbR*) to SCLAV-4217. The genes in the cluster are listed in Table 2.2.

Table 2.2. Genes from the cephamycin C biosynthetic cluster.

Type	Genes
Biosynthetic	*pcbAB* (early steps) *pcbC* (early steps) *cefD* (intermediate steps) *cefE* (intermediate steps) *cefF* (intermediate steps) *cmcI* (late steps) *cmcJ* (late steps) *cmcH* (late steps)
α-AAA formation	*lat* *pcd*
β-lactam resistance	*bla* *pcbR* *pbp74* *blp* *pcbR*
Transport	*cmcT*
Regulatory	*ccaR*
Other	SCLAV_4203 SCLAV_4209 SCLAV_4217

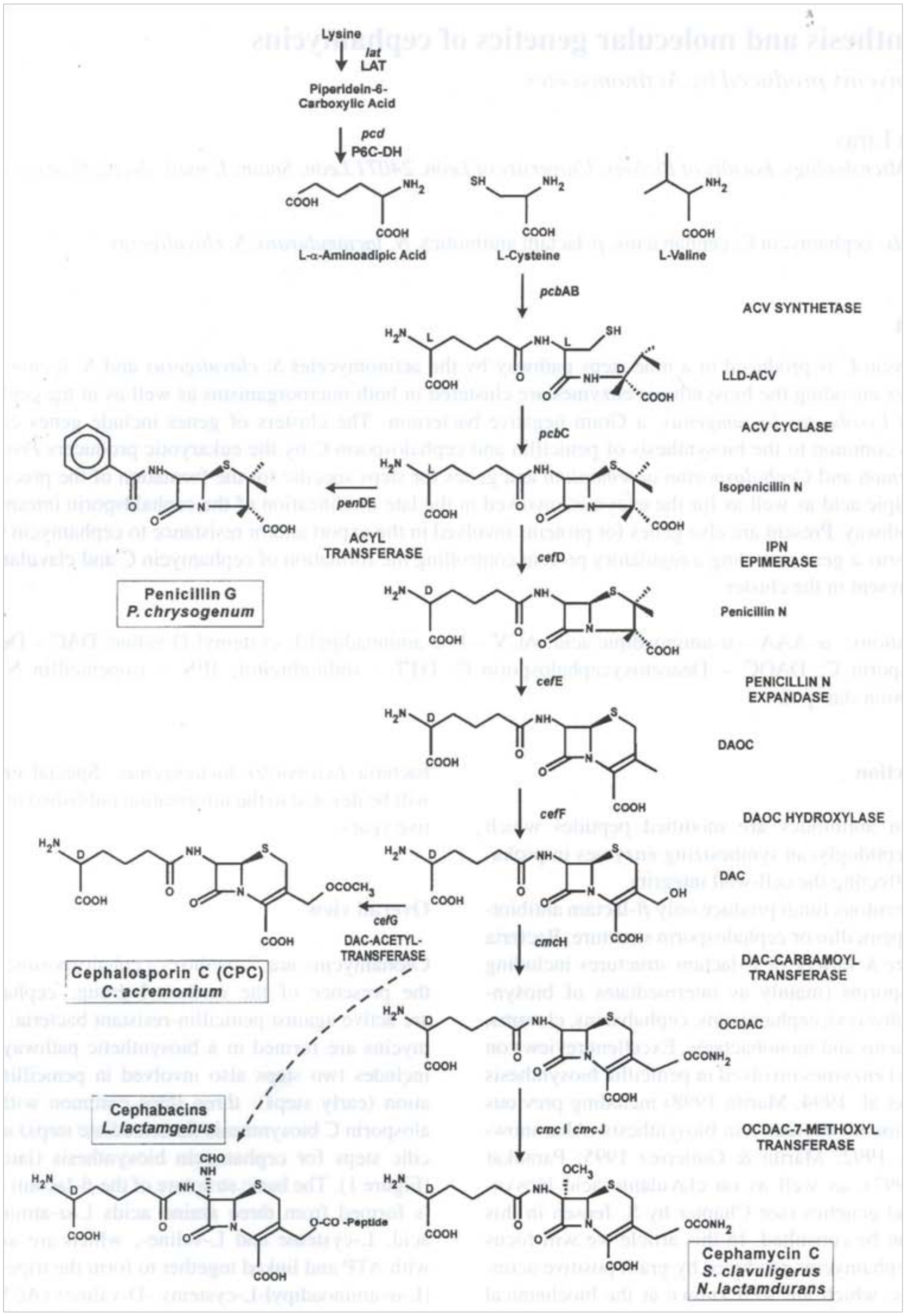


Figure 2.7. Cephamycin C biosynthetic pathway, from (Liras 1999). Other β-lactam antibiotics, like cephalosporin C, are also shown.

2.5.2.2 Clavulanic acid

Resistance of bacteria to β-lactam antibiotics occurs in bacteria that produce the enzyme β-lactamase, which inactivates β-lactam molecules (Sutherland 1991). Clavulanic acid is a potent β-lactamase inhibitor used in combination with amoxycillin and ticarcillin in the commercial products Augmentin^{TM} and Timentin^{TM}. The clavulanic acid–amoxycillin combination is used to treat urinary tract infections and lower respiratory tract infections (Brogden, Carmine et al. 1981). Clavulanic acid was the first clinically useful β-lactamase inhibitor. It was first described in 1976 by (Brown, Butterworth et al. 1976). Even though the combination clavulanic acid-amoxycillin has been in use for more than 20 years, it retains excellent activity against targeted pathogens (Saudagar, Survase et al. 2008). Clavulanic acid was isolated from *Streptomyces clavuligerus*, which was previously selected because of its ability to produce cephamycin C. Although produced by other organisms, its production by *Streptomyces clavuligerus* is the best studied one. Its β-lactamase inhibitor activity is associated with its *3R*, *5R* stereochemistry, as none of the other clavam metabolites (with *3S*, *5S* stereochemistry) exhibit inhibitor activity.

Clavulanic acid is produced in an eight step biosynthetic pathway from the precursors arginine and glyceraldehyde-3-phosphate (Liras and RodrÃ-guez-Garcia 2000). Among the intermediates are proclavaminic acid and clavaminic acid, which is a branching point for the production of either clavulanic acid or other clavams (Figure 2.8). The reaction(s) to convert clavaminic acid to clavulanic acid require stereochemistry inversion and are still unclear.

The cluster for clavulanic acid production is next to that of cephamycin C. The clavulanic acid biosynthetic cluster contains *claR*, a gene encoding a transcriptional regulator which regulates the late steps in the clavulanic acid pathway (Paradkar, Aidoo et al. 1998). Clavulanic acid production is also regulated by the SARP member CcaR, which is located in the cephamycin C biosynthetic cluster. The total length of the clavulanic acid biosynthetic cluster is 28 Kb, from the locus SCLAV_4178 to SCLAV_4197 (*ceaS2*).

Table 2.3 summarizes the genes in this cluster.

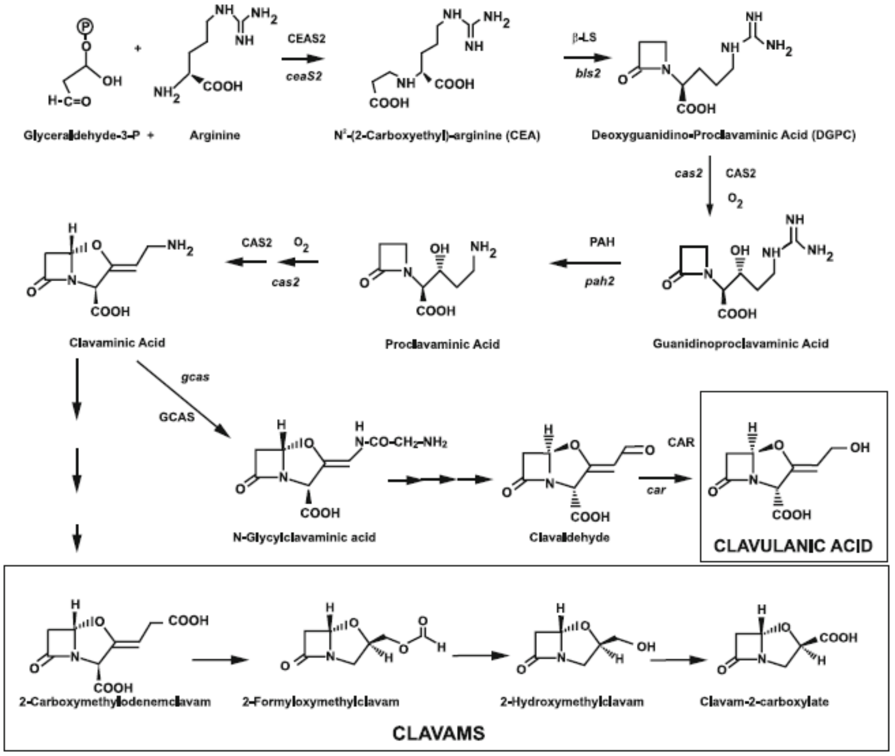


Figure 2.8. Pathway for clavulanic acid production, from (Liras, Gomez-Escribano et al. 2008).

Table 2.3. Genes from the clavulanic acid biosynthetic cluster.

Type	Genes
Biosynthetic	*ceaS2* *bls2* *cas2* *pah2* *gcaS* *car*
Penicillin-binding	*pbpA* *pbp2*
β-lactamase	SCLAV_4187
Regulatory	*claR*
Other	*fd* *cyp* *oppA1* *oppA2* *oat2* SCLAV_4178 SCLAV_4182 SCLAV_4184 SCLAV_4185 SCLAV_4186

2.5.2.3 Other clavams

In addition to clavulanic acid, *Streptomyces clavuligerus* produces several compounds with clavam structure but with *3S*, *5S* stereochemistry. These compounds include clavam-2-carboxylate, 2-formyloxymethylclavam, 2-hydroxymethylclavam, hydroxyethylclavam and alanyl-clavam (Liras, Gomez-Escribano et al. 2008). None of these clavams has β-lactamase inhibitory activity.

Three clusters are involved in the production of clavulanic acid and clavam compounds. The first is the previously described cluster for clavulanic acid, the second is the clavam gene cluster, which contains a clavaminate synthase isoenzyme, *cas1*. The third cluster is the paralogous gene cluster and contains several genes duplicated from the original clavulanic acid cluster and some genes which are paralogous from genes in the clavam gene cluster. Disruption of genes in the clavam and paralogous clusters has not been found to affect clavulanic acid production.

2.5.3 CcaR, a transcriptional regulator affecting cephamycin C and clavulanic acid production

CcaR was identified in 1997 by (Perez-Llarena, Liras et al. 1997) as a regulatory gene required for the production of both β-lactam antibiotics: cephamycin C and clavulanic acid (thus its name). *ccaR* disruptions abolish not only clavulanic acid production, but also cephamycin C production, and other clavams (Jensen and Paradkar 1999). The sequence of *ccaR* contains the rare codon UUA, which as in *Streptomyces coelicolor* is translated efficiently only by the tRNA encoded by *bldA*. In *Streptomyces clavuligerus* however, there is no effect on translation of *ccaR* on *bldA* mutants (Trepanier, Jensen et al. 2002). Expression of *ccaR* though is affected by another *bld* gene, in this case *bldG*. BldG deletion mutants do not produce clavulanic acid, cephamycin C, nor clavams (Bignell, Tahlan et al. 2005). Although ClaR is presumably involved in the conversion from clavaminic acid to clavulanic acid its exact role is not known yet.

CcaR belongs to SARPs, which are known to bind tandem heptameric sequences. The binding of CcaR has been confirmed to the intergenic region of the cephamycin C cluster genes *cefD–cmcI* and to its own promoter region, thus CcaR is an autoregulator (Santamarta, Rodriguez-Garcia et al. 2002). In the same study, binding of CcaR to the promoter regions of the cephamycin C cluster genes *lat* and *blp* was not detected. Although binding of CcaR to *lat* has been reported using *E. coli* crude recombinant CcaR (Kyung, Hu et al. 2001). Binding of CcaR to the promoter regions of the clavulanic acid cluster genes *claR* and *car* was also not detected by Santamarta et al. Thus the mechanism through which CcaR exerts control in the production of clavulanic acid has still not been deciphered.

A sequence for binding butyrolactone receptor proteins, ARE box was detected upstream of the *ccaR* translation start (position -890 bp). Brp, a butyrolactone receptor originally described by (Kim, Lee et al. 2004) binds to the ARE$_{\textit{ccaR}}$ box (Santamarta, Perez-Redondo et al. 2005). Brp is a repressor of cephamycin C and clavulanic acid. A second protein was reported to bind the ARE$_{\textit{ccaR}}$ box (Santamarta, Lopez-Garcia et al. 2007). However, pure recombinant AreB did not bind the ARE$_{\textit{ccaR}}$ box, unless small molecular weight extracts were added to the binding reaction.

Chapter 3
Tools for Genomic Studies

3.1 Summary

In this chapter an overview of next generation sequencing (NGS) technologies is presented. Particular emphasis is given to the second generation sequencing technologies by 454 and Solexa, as those were the technologies used in this work. A brief explanation of bioinformatics programs used for handling sequencing data is also included. The chapter closes with an overview of comparative genomics.

3.2 Next generation sequencing (NGS) technologies

Whereas previously researchers studied genes one by one, isolated from the rest of the genome, now genes can be studied as a whole. Genomics is the study not only of genes (structural genomics), but also of its function (functional genomics) (Hocquette 2005). Genomics and other “omics” fields, such as transcriptomics and proteomics, owe its development to sequencing (Hall 2007).

Sequencing has benefited greatly by advances in technology and has also motivated technology development. The development of microarray could not have been possible without genomic information. Microarrays, first used in 1995 by (Schena, Shalon et al. 1995) were the first technology to monitor the expression level of all genes simultaneously. The design of expression microarrays does not necessarily require a whole genome, but its full potential is realized only when thousands of coding sequences are analyzed at a time.

Sequencing has changed drastically since the first bacterial genome was sequenced in 1995 by (Fleischmann, Adams et al. 1995). At first, sequencing costs were elevated, and a clear bias towards sequencing human pathogens and model organisms was evident. Sequencing progressed slowly until the first NGS technology, that developed by 454 Life Sciences (Roche, Branford, CT), reached the market. Their parallel pyrosequencing method greatly reduced costs and allowed the direct sequencing of bacterial genomes, i.e., no cloning into plasmids was required. Solexa (Illumina, San

Diego, CA) and SOLiD (Applied Biosystems, Carlsbad, CA) soon followed with their NGS technologies.

Sequencing is now entering a new stage in which it is used not only to obtain the genome of the organism of interest, but also to characterize small RNAs, to measure gene expression levels, to study DNA methylation patterns, and to survey the diversity of organisms in environmental samples (Ansorge 2009). Sequencing is also opening the way for personal genomics, not only to understand human diversity, but also to detect mutations leading to illness.

A third generation of sequencing technologies is emerging. In addition to parallelization, these new technologies focus on single molecule sequencing in real time. In most cases single molecule sequencing will be achieved using sequencing by synthesis without amplification. Nanopores are also being investigated for use in DNA sequencing. Some of the technologies in development could be applied not only to DNA, but to RNA, and proteins. As sequencing costs continue to decrease and new technologies increase the output of sequencing methods, the most significant challenges will be in functional genomics, rather than in structural genomics. The blueprint obtained from genome sequencing is not enough to reveal gene function, additional analyses are necessary to make the transition from genome structure to gene function (Werner 2010). Unfortunately, functional genomics has not progressed at the same speed as structural genomics.

3.2.1 Parallel pyrosequencing

Parallel pyrosequencing (http://www.454.com/) was the first second-generation DNA sequencing technology to reach the market. It was introduced by 454 Life Sciences (Roche, Branford, CT) in 2005. In this technology, short stretches of DNA are amplified in an emulsion polymerase chain reaction (PCR) to create clonal clusters. The clusters are then sequenced by synthesis in a massively parallel way. When a nucleotide is incorporated into the nascent chain of DNA it is detected by the release of light, which is recorded by a camera.

In more detail (Figure 3.1), DNA is sheared and adaptors added. The single-stranded DNA fragments are then mixed with primer-coated 28 μm diameter beads, in a water-in-oil mixture that favors a ratio of one fragment per bead. Clonal amplification occurs in each of these beads when nucleotides and polymerase are added to the emulsion. Beads are deposited into a PicoTiterPlate, one bead per well. Sequencing then occurs in a parallelized way. Nucleotides are sequentially flowed into the plate containing hundreds of thousands of beads. When a complementary nucleotide is incorporated by the polymerase, a diphosphate group is released. The diphosphate group forms ATP, which is then used by a luciferase enzyme to emit light. Light emissions are recorded by a CCD camera. Apyrase is then flowed to stop the luciferase reaction, by degrading ATP. Unincorporated nucleotides are washed away and a new sequencing cycle begins. The intensity of the light signal is proportional to the number of nucleotides incorporated by the polymerase. Stretches of homopolymers can thus cause difficulties in this technology, if the light signal reaches saturation (Margulies, Egholm et al. 2005).

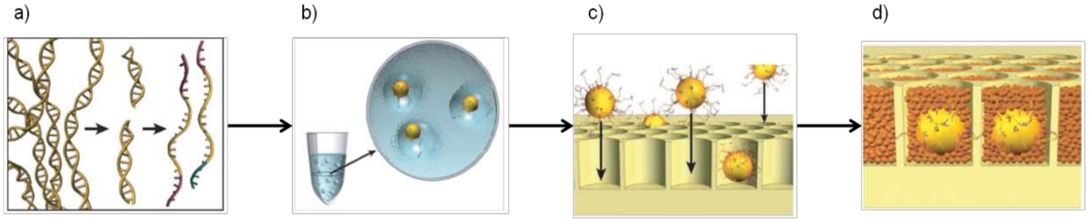


Figure 3.1. Schematic representation of parallelized pyrosequencing by 454. a) DNA is sheared and adaptors attached to it. b) The DNA is attached to beads in a ratio favoring one fragment per bead. The beads are trapped in drops in which PCR emulsion occurs. c) The beads with the clonal fragments are deposited in a plate. d) Enzymes required for DNA synthesis are deposited in the wells in small beads. Figure modified from (Margulies, Egholm et al. 2005).

3.2.2 Sequencing by reversible termination

In this technology, commercialized by Solexa (Illumina, San Diego, CA), DNA fragments are attached to a glass surface using adapters. Amplification is then performed to generate clonal clusters. The DNA in each cluster is made single-stranded, and sequencing primers are added and allowed to hybridize. These sequencing primers will be the starting point to the second strand of DNA that is generated by synthesis. Sequencing is done one nucleotide at a time. In each cycle all four nucleotides are added at once. Incorporation of a single nucleotide terminates growth of the DNA chain, as the nucleotides have been modified to contain reversible terminators. After the incorporated nucleotides have been identified using the attached fluorescent label, the fluorescent tag is removed and a 3' hydroxyl group is regenerated.

The hydroxyl group will be used to extend the growing DNA in the next cycle of nucleotide extension (Bentley, Balasubramanian et al. 2008) (Figure 3.2). The number of extension cycles has improved dramatically. When this technology was first released reads of only 35 bases were obtained with it. At the moment reads of 100+ bases can be achieved with this technology (http://www.illumina.com/).

In the case of paired-end sequencing, DNA is sheared to a specified size, with a narrow range. The sheared single stranded DNA is used as template for sequencing first, then used to create a second strand that will serve as the new template for sequencing. Sequencing primers are used to generate a second set of reads, that is separated from the original set of reads by a known length. Long insert sizes (longer than 1000 base pairs) are achieved by circularizing DNA, and randomly fragmenting it. The fragments containing the joining junction are selected as starting material in the sequencing process described above.

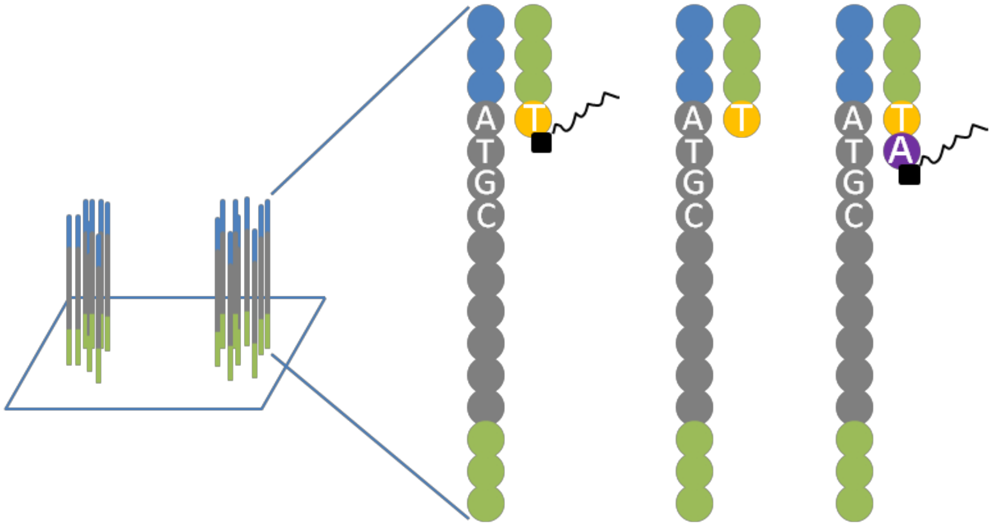


Figure 3.2. Solexa sequencing technology creates clonal clusters from the DNA fragments attached to a glass surface. Sequencing by synthesis starts from the primers and uses nucleotides containing a fluorescent label and a reversible terminator. The label is cleaved and a hydroxyl group regenerated. This allows the incorporation of the next nucleotide.

3.2.3 Other second generation sequencing technologies

In addition to 454 and Solexa, SOLiD is a second generation sequencing technology in current use. SOLiD, which stands for Sequencing by Oligonucleotide Ligation and Detection, was introduced in 2007 by Applied Biosystems As in the 454 technology, clonal amplification occurs by emulsion PCR, but the beads are 1 μm in diameter (Rothberg and Leamon 2008), much smaller than those used by 454. The beads are attached to a glass surface, generating a very high density random array. Sequencing occurs by ligation, instead of by synthesis.

In the SOLiD technology, a primer is hybridized to the adaptor and partially degenerate fluorescently labeled octamers are added. The identity of the central two nucleotides is known. The octamer gets one of four fluorescent labels according to the identity of the central dinucleotide. Octamers are allowed to hybridize to the template and are then ligated. At this point the slide is imaged to detect the four different fluorescent labels. The label is then removed by cleaving the fluorescent tag (three nucleotides), leaving five hybridized nucleotides, of which the sequence of two of them is known. This process is repeated, until a read of 35-50 base pairs (bp) is reached (MacLean, Jones et al. 2009). The sequence of every fourth and fifth nucleotide is known for the resulting reads. The newly synthesized strand is denatured from the template and removed. A new primer, which is off by one base pair with respect to the previous primer, is attached and the ligation cycle is restarted. The cycle of attaching a primer and ligation is repeated five times. Since the sequence is determined from a two base pair encoding system, the results are highly accurate. Because there are sixteen possible dinucleotide combinations and only four fluorescent labels are used, the identity of the first nucleotide has to be known to decipher the two-base encoding system, thus one base of the adaptor has to be sequenced (Turner, Keane et al. 2009). Knowing the first nucleotide and the color of the fluorescent tag, the identity of the second nucleotide is recovered. This nucleotide becomes then the first nucleotide for the next dinucleotide (Shendure and Ji 2008) decoding. This results in reads with > 99.94% base calling accuracy (http://www3.appliedbiosystems.com/). Because each instrument contains two independent flow cells (Mardis 2008) that can be divided in octets and 20 tags are available for multiplexing, up to 320 samples can be sequenced in a single run.

3.2.4 Single molecule sequencing

Several technologies that are in development will allow the direct sequencing of single molecules of either DNA or RNA without amplification. In most cases sequencing will also occur in real time, that is, at a speed close to that of DNA synthesis within a living cell. Among the technologies that are in development, one of the most promising ones and that seems to be ahead in becoming a commercial platform is that by Pacific Biosciences.

Pacific Biosciences sequencing technology (http://www.pacificbiosciences.com/) is based on Single-Molecule Real-Time, SMRT^{TM} (Eid, Fehr et al. 2009). The SMRT^{TM} chips consist of thousands of zero-mode waveguides (ZMWs) fabricated on a 100 nm metal film deposited on a silicon dioxide substrate. Each ZMW is a cavity with a 10–50 nm diameter in which a DNA polymerase is immobilized. A single-stranded template is used, and DNA synthesis, occurring at a speed of tens of nucleotides per second is directly observed. ZMWs become visualization chambers with a detection volume of 20 x 10^{-21} liters (20 zeptoliters). The minimal detection volume allows the continuous observation of single molecules against very low background levels. DNA synthesis occurs when nucleotides phospholabeled with different fluorophores are flowed at very high concentrations. When a nucleotide is incorporated by the DNA polymerase, it remains in the detection volume tens of milliseconds, in sharp contrast to the case when nucleotides simply flow in and out of the detection volume, which lasts only a few microseconds. The rest of unincorporated nucleotides float outside of the detection volume, in the dark, without lighting up. Fluorophores are cleaved as the DNA strand grows, and diffuse out of the detection volume. This results in the production of natural strands of DNA, without incorporated labels that cause steric hinderance or that contribute to background noise. This process does not require washing, scanning, or sequential flowing of nucleotides. The action of the DNA polymerase does not have to be stopped with every cycle. Detection occurs continuously and simultaneously across all the thousands of ZMWs. This technology has the potential of generating accurate reads longer than those generated by Sanger technology. Insertion and deletion errors in sequencing are directly caused by the DNA polymerase, instead of by multiple reactions going out of synchrony. Read length is a function of the enzyme processivity, so as enhanced polymerases are developed, read length will increase further. Future applications will extend to sequencing RNA directly, by using a reverse transcriptase instead of a DNA polymerase, and observing protein synthesis directly, by using ribosomes and labeled tRNAs.

The use of nanopores for sequencing purposes is also being investigated by several groups and companies. Two types of nanopores are under study for use in sequencing:

biological and solid state nanopores. Biological nanopores are formed by organic molecules, like proteins. The most common protein in use so far for biological nanopores is α-hemolysin. Because of their narrow diameter, α-hemolysin nanopores allow the pass of single stranded DNA only. The potential use of biological nanopores was demonstrated in 1996 by Kasianowicz et al. (Kasianowicz, Brandin et al. 1996). Solid state nanopores are made using nanofabrication techniques. Solid state nanopores allow the pass of double stranded DNA (Dekker 2007). Whichever the type of nanopore though, the identification of molecules is done as they pass through the nanopore, by measuring the disturbances in electric currents and forces. In addition to DNA sequencing, nanopores could also be used for RNA sequencing, and protein identification.

3.3 Read alignment

The output of NGS technologies is several orders of magnitude higher compared to that of traditional Sanger sequencing. However, the read length is shorter and error rate is higher. Their use for genome and transcriptome sequencing is thus more challenging (De Bona, Ossowski et al. 2008). Alignment of short reads is done not only for comparison purposes to reference genomes, but also to find out which reads were assembled into contigs, as many assemblers do not keep read tracking (Li and Homer 2010). The need to align short sequences to longer ones also arises in the case of identification of motifs, restriction enzyme sites, and miRNA sequences (Prufer, Stenzel et al. 2008).

Alignment algorithms use indices to handle the alignment information. The indices can be for the reads, for the reference sequence, or for both (Li and Homer 2010). Based on the type of index, alignment algorithms can be categorized as those using hash tables, and those using suffix trees. Examples of programs hashing the reads and scanning the genome include MAQ (Li, Ruan et al. 2008), SeqMap (Jiang and Wong 2008), RMAP (Smith, Xuan et al. 2008), and SHRiMP (Rumble, Lacroute et al. 2009). Programs constructing hash tables for the reference sequence include SOAP v.1 (Li, Li et al. 2008), MOM (Eaves and Gao 2009), and BFAST (Homer, Merriman et al. 2009).

Algorithms based on suffix trees identify exact matches and build inexact alignments sustained by the exact matches. A suffix tree is a data storage structure, which in the case of sequence alignment allows the efficient finding of subsequences. Examples of programs using suffix trees include SOAP v.2 (Li, Yu et al. 2009), BWA (Li and Durbin 2009), and Bowtie (Langmead, Trapnell et al. 2009). These programs use the Burrows-Wheeler transform (BWT) (Burrows and Wheeler 1994) algorithm. BWT was developed for data compression and has found application in allowing the efficient search of large texts.

3.3.1 Variant calling

Determination of genetic variation in bacteria can be the result of comparing a newly sequenced organism to a previously sequenced reference or of comparing two strains with different phenotypes. Whereas in the first case we expect a high number of differences, in the second case the differences should be few (Nusbaum, Ohsumi et al. 2009). Information on DNA polymorphisms characterized by DNA sequencing can be used to distinguish between bacterial strains (Li, Raoult et al. 2009). The identification of these differences is generally done by aligning sequencing reads. The differences can be of any size, and correspond to deletions, insertions, duplication events, or single nucleotide differences. Single nucleotide differences are of no less importance than bigger deletions or insertions, as they can be responsible for dramatically different phenotypes, for example antibiotic resistance.

3.3.2 Bowtie

Bowtie (Langmead, Trapnell et al. 2009) is a memory-efficient program for aligning short reads to large genomes that can be run on multiple processor cores to achieve high alignment speeds. Bowtie can handle short and long reads (4-1024 bp) and the input can be reads with a mixture of lengths. Bowtie (with the "–best" option) reports the best alignment based on minimizing mismatches in the seed portion of the read. The algorithm uses BWT to index the reference genome. The indexes generated by Bowtie have a small memory footprint. Bowtie samples short substrings of the reference and compares the query and the sampled substrings allowing only a few differences. It requires alignment of the full read, which can have disadvantages when aligning to a genome containing scaffolds.

3.3.3 SAMtools

The Sequence Alignment/Map format is becoming the default format for storing read alignments. It has the advantage of being flexible and compact and allows efficient access to the information contained in the alignment. SAMtools (Li, Handsaker et al. 2009) are a series of utilities dealing with post alignment information. The alignment information should be stored in the SAM format. Among the utilities included are those for sorting, merging, and indexing alignments. SAMtools can also perform variant calling, and has a basic alignment viewer. SAMtools can also perform conversion from SAM to other formats, like Binary Alignment/Map, BAM, which is basically a SAM file in binary form. Most alignment programs can now generate output in SAM format. The sorted and indexed information contained in SAM files can be used by applications to work on particular regions of the genome, without having to load the complete alignment file into memory.

3.4 *De novo* assembly and scaffolding

De novo assembly is the reconstruction of a genome without assistance from any previous information from genomes, transcripts, or proteins. In the case of assemblies using NGS data, the generated reads are short but in high abundance, for most projects in the millions of reads. The oversampled template DNA can then be assembled because the reads overlap. This overlap is determined by sequence alignment (Miller, Koren et al. 2010).

The output of an assembly is contigs and sometimes scaffolds or supercontigs. Contigs are generated by the consensus sequence of overlapping reads. Supercontigs will include contigs and stretches of N's. The N's represent gaps in the sequence, for which the identity of the position could not be determined, but the estimated length of the gap is inferred. Assemblers use information from paired-end reads and their estimated insert size to orient and separate the contigs with respect to each other. Paired-end reads can also help in resolving repeats.

Among the types of information that could be taken into account during assembly are the quality scores of each base in the reads, and the error profile for the particular technology used to generate the data. The error profile differs for the different technologies; short deletions for example are the most common type of error in the 454 technology, whereas for Solexa it is substitutions. Short deletions or insertions can occur in 454 by reactions going out of synchrony, and by stretches of homopolymers for example. In Solexa, each nucleotide added terminates the reaction and this offers the advantage that only one nucleotide is incorporated.

Assemblies can be assessed by their statistics. For example the number of generated contigs, their total length, the number of reads assembled into contigs, and other contig statistics like average and maximum length. A common number to report is the N50. N50 is calculated by ordering the contigs from longest to shortest and finding the contig up to which the combined lengths represent half of the total length. Accuracy, however, could only be assessed if a high quality reference genome sequence is available.

Algorithms for NGS data assembly are based on graphs, which is a set of nodes and edges between the nodes. Assemblers can be classified based on the type of graph used as either Overlap/Layout/Consensus or de Bruijn Graphs. The most popular assemblers, including Velvet and SOAPdenovo, use *k*-mer graphs, which are de Bruijn graphs. A *k*-mer is a sequence of *k* consecutive nucleotides which is used to represent reads. A *k*-mer graph contains nodes representing subsequences and edges represent overlaps. Algorithms using Overlap/Layout/Consensus involve all vs. all pair-wise read comparisons and multiple sequence alignment. de Bruijn graph-based algorithms do not require all vs. all read comparisons and are thus preferred when large number of reads are available.

3.4.1 GS De novo assembler (Newbler)

Newbler is an Overlap/Layout/Consensus assembler developed specifically for data generated by the 454 platform. It is distributed with the 454 sequencing machines and the source code is only available under certain conditions. Originally developed for reads in the 100 bp range generated by the GS 20 machine, it has been revised several times for longer reads (which are at this moment reaching 400 bp routinely) and to include the use of paired-end information. Newbler is a two step algorithm in which unitigs are generated and then used to create larger contigs. Unitigs are high confidence contigs (http://sourceforge.net/apps/mediawiki/wgs-assembler/index.php?title=Celera_Assembler_Terminology). Newbler takes advantage of instrument metrics to resolve problems like number of bases in homopolymers and other base calling errors.

3.4.2 Velvet

Velvet (Zerbino and Birney 2008; Zerbino, McEwen et al. 2009) is a de Bruijn graph-based assembler. The four stages in this algorithm are:

1. Hash reads into *k*-mers
2. Construct the graph
3. Correct errors
4. Resolve repeats

The two last steps are critical. The elimination of error merges sequences that should be together, whereas the repeat solver separates paths sharing local overlaps. The paths represent the reads traversing a *k*-mer graph. Paired-end read information is used to resolve tangles created in the de-Bruijn graph by sequence repeats and to construct scaffolds. The second step, the construction of the graph, is the most time and memory consuming step. Velvet can handle paired-end reads with different insert sizes. Among the parameters that can be adjusted by the user are *k*-mer size, the minimum frequency of expected *k*-mers and the expected coverage of the genome. The *k*-mer size must be odd, to avoid the possibility of a *k*-mer being its own reverse complement.

3.4.3 SOAPdenovo

SOAPdenovo (Li, Zhu et al. 2010), as Velvet, is based on de Bruijn graphs. It has been used successfully to assemble mammalian genomes (human and panda) (Li, Fan et al. 2010). This algorithm filters and corrects the reads based on thresholds for *k*-mer frequencies (Miller, Koren et al. 2010). Nodes represent *k*-mers and edges represent read paths. This program resolves tangles in the path by using information from coverage to determine which path to keep. Contrary to Velvet, it does not track which reads are assembled into which contigs. For constructing scaffolds SOAPdenovo maps paired-end reads to the contigs it assembled. When multiple libraries are available they are used sequentially, from the one having the shortest insert size to the one having the longest.

3.4.4 Other assemblers

The first programs capable of handling the short reads with uniform length of the Solexa platform were available in 2007. The read length at that point was so short (25 bp) that assemblers performed iterative extensions. Contigs were extended by using reads that overlap the contig ends and for which the overlap was above a certain length. The first assembler of this type was SSAKE (Warren, Sutton et al. 2007), and was soon followed by SHARCGS (Dohm, Lottaz et al. 2007). SHARCGS performs three separate assemblies and then attempts to join the contigs generated by the three sets. VCAKE (Jeck, Reinhardt et al. 2007) allowed the incorporation of reads with inexact overlaps.

The Overlap/Layout/Consensus algorithm has been implemented for short, uniform length type of data in two programs: Edena (Hernandez, Francois et al. 2008) and Shorty (Hossain, Azimi et al. 2009). Edena first filters out all duplicate reads, then constructs contigs based on perfect overlaps. Shorty uses a few long reads as seeds and expands them with short reads. de Bruijn graphs were first used in the EULER assembler (Pevzner, Tang et al. 2001). ALLPATHS (Butler, MacCallum et al. 2008; Maccallum, Przybylski et al. 2009) is based on de Bruijn graphs too. ALLPATHS requires paired-end reads with two insert sizes. Although we were able to run ALLPATHS successfully using example data, we were unsuccessful using our real data, which nonetheless corresponds exactly to the type of input required by ALLPATHS.

3.4.5 MUMmer

MUMmer is an algorithm for the rapid alignment of whole genomes. It is based on suffix trees. Alignments are based on shared maximal unique matches (MUMs) between subsequences of both queries (Delcher, Kasif et al. 1999).

In addition to its use for comparing whole finished genomes, MUMmer can be used to compare contigs from draft assemblies to finished genomes. Two sets of contigs, either from two different organisms or from two stages of the same project can also be efficiently aligned using MUMmer.

Several upgrades to MUMmer have been released over the years. The second version (Delcher, Phillippy et al. 2002) only constructs the suffix tree for the reference genome and scans the query genome with it. Thus the MUMs are only unique in the case of the reference but not necessarily in the query. This reduces memory requirements, as only one suffix tree is stored in memory. The second version also introduced the modules NUCmer and PROmer, for aligning nucleotides and proteins, respectively. When NUCmer is used to align a set of contigs to a reference, the output is the order and orientation of the contigs with respect to that reference. PROmer translates DNA to amino acids and then compares the alignments. These alignments can reveal much older relationships that at the DNA level present minimal conservation. MUMmer has also been modified to handle mapping of reads to reference genomes (Schatz, Trapnell et al. 2007).

3.5 Open reading frame (ORF) prediction

Prediction of open reading frames (ORFs) in bacteria, although simpler than in eukaryotes, is nonetheless a challenging task. ORFs determination can be similarity-based or *ab initio*. Similarity-based approaches infer ORFs from similarity searches, especially to close organisms. *Ab initio* approaches extract all the information from the sequence using statistical techniques (Do and Choi 2006). *Ab initio* approaches have to make decisions for example when overlapping ORFs are found, and they have to identify the translation start among several possible start codons. Some algorithms for ORF prediction use additional information such as identification of promoters, terminators, and regulatory motifs, to help improve their predictions.

Among the *ab initio* gene prediction programs are GeneMark, Glimmer (Salzberg, Delcher et al. 1998; Delcher, Harmon et al. 1999), GeneLook (Nishi, Ikemura et al. 2005), ZCURVE (Guo, Ou et al. 2003), and Prodigal (Hyatt, Chen et al. 2010). Genbank lists ORF predictions for all publicly available bacterial genomes using GeneMark, GeneMark.hmm, Glimmer, and Prodigal.

3.5.1 Glimmer

Glimmer (Salzberg, Delcher et al. 1998; Delcher, Harmon et al. 1999) is a program for the identification of microbial genes which uses interpolated Markov models (IMMs). In the case of fixed order k^{th} Markov models, a nucleotide is predicted based on the previous *k* nucleotides. However this requires that all possible *k*-mers are represented in the data multiple times. Since this is not always the case, the IMM uses only the *k*-mers for which there is enough data. The *k*-mers are used in a weighted manner, so that the *k*-mers present frequently have higher weights. Glimmer uses *k*-mer sizes from one to eight. Glimmer uses the IMM to identify putative genes and then scores them in all six reading frames. The output is a list of putative gene coordinates.

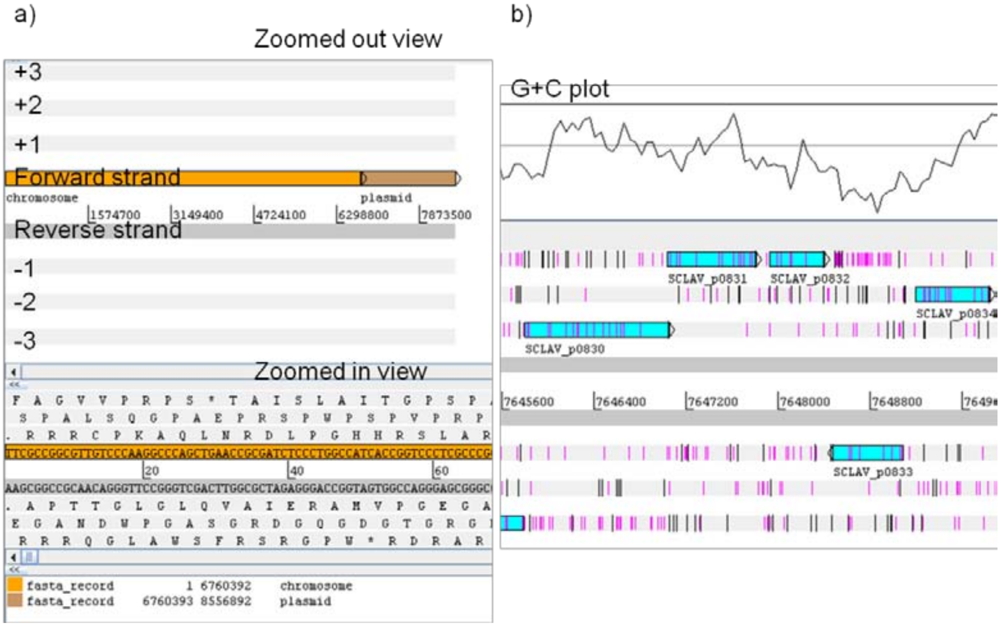
3.5.2 GeneMark

GeneMark is a family of programs for the prediction of genes in prokaryotes, eukaryotes, and viruses. The family is integrated by three programs: GeneMark, GeneMark.hmm, and GeneMarkS. GeneMark uses Markov models to describe nucleotide subsequences that have different rules for the case of coding and non-coding DNA regions (Borodovsky and McIninch 1993). GeneMark.hmm (Lukashin and Borodovsky 1998) uses hidden Markov models to predict gene boundaries, which are modeled as transitions between hidden states. GeneMark.hmm also uses information on the ribosome binding site pattern to improve the determination of the start codon. GeneMarkS (Besemer, Lomsadze et al. 2001) is an iterative method which focuses in the identification of genes' starts. Basically GeneMarkS runs GeneMark.hmm iteratively and uses multiple sequence alignment to learn a new model regarding the upstream region of the genes. GeneMark.hmm is run until convergence. GeneMarkS is a self-training program, which in contrast to GeneMark and GeneMark.hmm does not require pre-computed species models. The output includes the gene boundaries, gene length, and the strand in which it localizes. Genes are also labeled as either “typical” or “atypical”, which refers to codon usage. “Typical” genes are those that have a codon usage similar to most of the genes in the species, whereas “atypical” have different codon usage and most likely represent laterally transferred genes (Besemer and Borodovsky 2005).

3.6 Visualization tools

3.6.1 Artemis

Artemis (Rutherford, Parkhill et al. 2000; Berriman and Rutherford 2003) is a Java-based program for the visualization of sequence information in the context of its six frame translation. It can handle information from a single gene and up to full genomes, especially compact ones, such as those of bacteria. Input information can be simple FASTA files for sequence information only or either EMBL or GenBank files for features such as contigs, gaps, coding sequences, gene names, etc. Start and stop codons can be easily displayed after the correct genetic code has been selected. Plots for G+C content can easily be displayed as well. Visualization can be from the nucleotide level and up to the whole genome, and information can be loaded in individual tracks.



```
F A G V V P R P S * T A I S L A I T G P S P
 S P A L S Q G P A E P R S P W P S P V P R P
. R R R C P K A Q L N R D L P G H H R S L A R
TTCGCCGGCGTTGTCCCAAGGCCCAGCTGAACCGCGATCTCCCTGGCCATCACCGGTCCCTCGCCCG
AAGCGGCCGCAACAGGGTTCCGGGTCGACTTGGCGCTAGAGGGACCGGTAGTGGCCAGGGAGCGGGC
. A P T T G L G L Q V A I E R A M V P G E G A
E G A N D W P G A S G R D G Q G D G T G R G
 R R R Q G L A W S F R S R G P W * R D R A R
```

🟧	fasta_record	1	6760392	chromosome
🟫	fasta_record	6760393	8556892	plasmid

Figure 3.3. Screenshots from Artemis showing some of its visualization features. a) Zoomed out view (top) showing two genomic elements. The zoomed in view (bottom) shows the nucleotide sequence. b) Some genes (blue arrows) in their corresponding translation frame. Black vertical lines indicate stop codons, and pink vertical lines indicate start codons. The plot on top corresponds to G+C content.

3.6.2 BamView

BamView (Carver, Bohme et al. 2010) is a Java-based program that allows the visualization of read alignments stored in the Binary Alignment/Map (BAM) format. BamView can not only handle read alignment information from resequencing projects, in which coverage is in the same order of magnitude along the genome, but also transcriptome sequencing data, in which coverage can vary by orders of magnitude.

BamView can be used in conjunction with the genome browser Artemis (Section 3.6.1 thus allowing the visualization of coverage directly on top of genomic features. In the case of paired-end reads, BamView can display the alignment information with respect to the inferred insert sizes. Bases that differ in the reads with respect to the reference can be displayed in red, thus when polymorphisms are present they are clearly visible as red vertical lines.

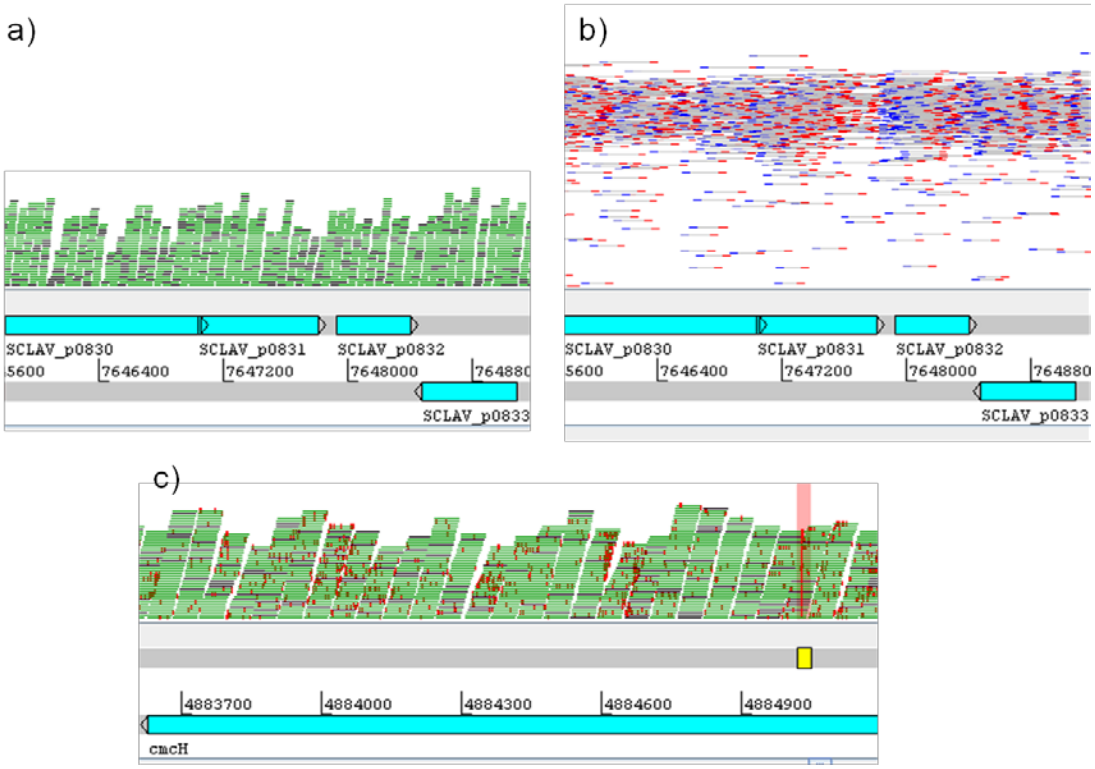


Figure 3.4. Screenshots of BamView within Artemis. a) Stacked view, each read is represented as a horizontal line, green if the read is duplicated and black if it is unique. b) Inferred size view. Paired-end reads are represented by blue and red (forward and reverse) horizontal rectangles joined by a gray line indicating the insert. c) Nucleotide differences are colored in red, thus a SNP appears as an easily identifiable red vertical line (highlighted in pink).

3.7 Comparative genomics

Genomics is the science of an organism complete genome (Rothschild and Plastow 2008), as the result of its sequencing. This includes the determination of characteristics of the genome, its non-coding regions and its genes. The basic characteristics of a genome include the number of chromosomes and plasmids and their structure (linear or circular), total length, G+C content, and ORF determination. ORF characteristics include their length, function, and similarity to other genes within the same genome and to other genomes. Comparison between several strains of the same species allows determination of single nucleotide polymorphisms (SNPs) and genome-wide mutations.

Comparative genomics focuses in finding the similarities and differences in the genome sequences of organisms and strains. Bacterial genome sequencing has played a vital role in comparative genomics. In 1999, the genome sequence of two isolates of the same species were available for the first time. The two isolates compared were of the bacteria *Helicobacter pylori* (Alm, Ling et al. 1999). Of the completed and ongoing sequencing projects, a high proportion corresponds to bacteria. Until August 2010, genomes online (http://www.genomesonline.org) reported finished genomes for 1136 bacteria, 133 eukaryotes, and 92 archaea. Ongoing projects are reported for 4,816 bacteria, 1548 eukaryotes, and 186 archaea. Eighty percent of this high number of bacterial genomes, however, is part of three phylum only: Proteobacteria, Firmicutes, and Actinobacteria. This puts in evidence that still most of the bacterial diversity remains unexplored. A couple of initiatives are trying to fill the gap in our understanding of bacterial complexity and function. One such initiative is the Genomic Encyclopedia of Bacteria and Archaea (GEBA), which was launched by JGI (http://www.jgi.doe.gov/programs/GEBA/).

Early on, a bias towards sequencing genomes based on their potential practical application was clear (Kyrpides 2009). Pathogenic bacteria and microbes used in industrial production were among the first bacterial genomes sequenced. The cost reduction in sequencing has now democratized sequencing (McPherson 2009), however there is still a gap to fill with respect to functional genomics.

Comparative genomics has shown that even genomes of the same species exhibit high genetic diversity (Binnewies, Motro et al. 2006). Bacterial genome size ranges from 0.144 million base pairs (Mb) for *Candidatus* Hodgkinia cicadicola (McCutcheon, McDonald et al. 2009)[4] to 13 Mb for *Sorangium cellulosum* (Schneiker, Perlova et al. 2007). G+C content in bacteria ranges from 16% in *Candidatus* Carsonella ruddii (Nakabachi, Yamashita et al. 2006) to 75% in *Anaeromyxobacter dehalogenans* (Thomas, Wagner et al. 2008). Even coding density, which was assumed to be high in all prokaryotes varies from 51% to 95% (Binnewies, Motro et al. 2006).

Strains of the same species also exhibit high genomic diversity. For example, the genome size of 33 strains of the model organism *E. coli* ranges from 4.6 Mb to 5.7 Mb. The comparison of the genomes of *E. coli* strains K12, O157:H7 and CFT073 found that only 39% of genes were common to all three strains (Welch, Burland et al. 2002; Medini, Serruto et al. 2008). Some of the mechanisms for generating diversity include mobile genetic elements such as plasmids, transposons, conjugative transposons, bacteriophages, integrons, and insertion sequence elements (ISs).

As close organisms are compared, the notion that bacterial genomes consist of a backbone genome, containing core genes, and adaptation modules, containing flexible genes, has strengthened. Genes for essential functions like replication, transcription, and translation are located in the backbone (Medini, Serruto et al. 2008). Adaptation modules may or may not be present in a particular strain, but never are all adaptation modules present in one strain. This has led to the concept of pangenome, which has been defined as all the different genes present in different genomes of the same species (Tettelin, Masignani et al. 2005). A pangenome contains one representative of sequences found in multiple strains in the group of genomes analyzed, and all the genes found uniquely in any of the strains being compared.

The tools for comparing bacterial genomes range from the simple length comparison to chromosome alignment. Synteny, for example, focuses on identifying chromosomal translocations and inversions. Other genome characteristics that are frequently compared include the G+C content, presence of repeats, and promoter analysis (Medini, Serruto et al. 2008). Comparisons can also focus on specific types of proteins, for example sigma factors, transcription factors, secreted proteins, membrane protein, or two-component systems.

However, it is obvious that sequencing and functional genomics have not progressed at the same rate. Functional genomics' aim is to assign functions to unknown genes, as simply knowing the sequence and location of genes is only the starting point in understanding biological systems (Holtorf, Guitton et al. 2002). Information obtained from transcriptomics, proteomics, metabolomics, and phenomics studies can help in suggesting gene functions. Although some of these areas perform parallelized analysis (especially transcriptomics), others still collect information at a more individual level (especially phenomics).

Chapter 4

Consensus Sequence-based Assessment of Genome-wide Regulatory Networks in Streptomyces coelicolor

4.1 Summary

Secondary metabolite production in *Streptomyces*, soil-dwelling bacteria, is regulated by complex mechanisms linked to morphological changes. The availability of a large number of transcriptome data from *Streptomyces coelicolor* offers incentives to explore these interacting regulatory mechanisms and develop biologically meaningful hypothesis. Genomic features and transcriptome data were previously used in our lab to construct a whole genome operon prediction for *Streptomyces coelicolor*. Transcriptome data at the cistron level was also previously used to infer regulatory networks in this model actinomycete. In this work, the network modules resulting from that approach are analyzed in terms of their biological coherence. Presence of experimentally verified interactions, functional enrichment, and consensus sequences in the upstream region of the elements belonging to the network modules were assessed and used as supporting evidence for the validity of the inferred network modules.

4.2 Introduction

Streptomyces are soil-living bacteria well known for their production of secondary metabolites, many of which are used as antibiotics, anti-tumor agents, immunosuppressants, and anti-cancer agents (Paradkar, Trefzer et al. 2003). *Streptomyces* present a complex life cycle that includes formation of aerial mycelia and spores, and changes in morphology and secondary metabolite production are known to be linked (Chater 1993).

The genome of *Streptomyces coelicolor*, the model organism for actinomycetes (Donadio, Sosio et al. 2002), was sequenced in 2002, and revealed one of the most complex bacterial genomes sequenced up to that day. The genome includes 7825 genes (Bentley, Chater et al. 2002), with a high percentage involved in regulatory roles. These regulatory roles can be pleiotropic or only at the pathway specific level. The genome sequence of *Streptomyces coelicolor* revealed more than twenty secondary metabolite clusters.

Deciphering the complex regulation of secondary metabolite production has been of great interest in the *Streptomyces* community. Extensive work has been done with knock-out and disruption mutants with the goal of understanding the role of individual genes in regulatory networks. The pathway specific regulators in the main antibiotic clusters (RED, ACT, and CDA) have been characterized. These pathway specific regulators include *redD*, and *redZ* for the undecylprodigiosin (RED) antibiotic; *actII-ORF4* for the actinorhodin (ACT) antibiotic; and *cdaR* for the calcium-dependent antibiotic (CDA). Additionally, other pleiotropic regulators have also been characterized, including *afsR* and multiple two-component systems (e.g., *phoRP*, and *absA1A2*).

Regulation, however, is a dynamic process, with interacting smaller networks. Membership into networks is not permanent. In the case of *Streptomyces*, the number of genes is more than that in lower eukaryotes and almost double the number of genes in *E. coli*. Thus the networks linking growth, morphology, and secondary metabolite production are expected to be complex. Furthermore, as for all bacteria, genes can be co-transcribed, and it is at this level of cistrons that regulation occurs, rather than at the level of individual genes.

Microarray data, especially time series microarray data, has been used for computational approaches to reconstruct gene networks. Due to the dynamic nature of regulation, data should include studies for a variety of mutants, or for strains under diverse conditions. By mining gene expression data over diverse conditions and time profiles, sets of cistrons (in bacteria) that present correlation can be identified.

There are two main approaches for network reconstruction methodologies: information theoretic methods and dynamic Bayesian networks (DBNs). Information theoretic methods consist of two steps. First, mutual information (MI) among the expression profiles is calculated for all possible gene (or cistron) pairs. MI is a measure of relatedness than can detect non-linear interactions. Second, MI values are compared in order to infer interactions. This second step is what differentiates the multiple algorithms. Some of the information theoretic methods include relevance networks (RELNET) (Butte and Kohane 2000; Butte, Tamayo et al. 2000), Context Likelihood of

Relatedness (CLR) (Faith, Hayete et al. 2007), Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) (Basso, Margolin et al. 2005; Margolin, Nemenman et al. 2006; Margolin, Wang et al. 2006), and Minimum Redundancy Networks (MRNET) (Meyer, Kontos et al. 2007). These algorithms are included in the R package minet (Mutual Information NETwork inference) (Meyer, Lafitte et al. 2008).

In the case of ARACNE, the algorithm used our work, indirect relationships are eliminated by using data process inequality (DPI), a characteristic of mutual information. The ability to discriminate between direct and indirect interactions is a desirable trait of network reconstruction algorithms, as the number of false positives will be reduced. Information theoretic algorithms, however, lack directionality.

4.3 Materials and methods

4.3.1 Operon prediction and network reconstruction

Operon prediction and network reconstruction with ARACNE were done previously in our lab by Dr. Salim Charaniya (Charaniya 2008). Briefly, microarray data for 524 cell samples was obtained from in-house generated data and the public repository databases Stanford Microarray Database, Gene Expression Omnibus (GEO), and Array Express. Features used for operon prediction included: conservation of gene order, functional similarity, intergenic distance, and gene expression similarity. SVM ^{light} was used to construct models for the prediction of operons. Recall, false positive rate and area under ROC curves were used to assess the performance of classifiers as described in (Charaniya, Mehra et al. 2007).

Transcriptional networks were predicted on the whole genome using ARACNE (Basso, Margolin et al. 2005; Margolin, Nemenman et al. 2006). As ARACNE requires a complete gene expression matrix, the *k*-nearest neighbor method was used to fill in any missing values. Data for genes with low expression dynamics or with a large number of absent flags was filtered out. A *p*-value of 1.0×10^{-9} was used as threshold for mutual information (MI) and a DPI tolerance of 0.05 was used as criteria to remove possible indirect interactions. Predicted networks were visualized in Cytoscape within ARACNE.

4.3.2 Consensus sequence determination

In addition to the functional enrichment previously calculated by Dr. Charaniya using Fisher's exact test (Charaniya 2008), the presence of consensus binding sequences in the upstream region (300 bp) of the first gene of cistrons belonging to the same network module was examined using MEME version 3.5.7 (Bailey and Gribskov 1998; Bailey, Williams et al. 2006). A zero order background Markov model was determined by calculating the fraction of each base in the upstream region of all 5346 predicted cistrons. The fractions were as follows: A:0.153, C:0.351, G: 0.347, and T:0.14. A threshold *E*-value (*E*-value$_{\text{threshold}}$) was determined by randomly shuffling the sequences for each network and assessing the presence of consensus binding sequences. This was repeated five times and the minimum *E*-value used as threshold for the case using the real sequences.

A list of publications associated with individual genes was obtained by querying the tables behind StrepDB (http://strepdb.streptomyces.org.uk/cgi-bin/dc3.pl?accession=AL645882&start=4291472&end=4302043&iorm=map&width=900) by SQL query with Perl scripts. Further parsing of data was done with scripts written in Perl.

4.4 Results

4.4.1 Genome-wide regulatory network of *Streptomyces coelicolor* reconstruction

Operon prediction and network reconstruction (using ARACNE) were done previously in our lab by Dr. Salim Charaniya (Charaniya 2008). Based on gene expression similarity, intergenic distance, conservation of gene order, and functional similarity a total of 5346 transcription units were predicted. Of these transcription units 3957 are monicistronic and 1389 are polycistronic. Polycistrons range from 2 to 21 genes. Transcriptome data at the cistron level was used to infer regulatory networks in ARACNE. In addition to transcriptome data, ARACNE requires a list of regulators. In our case the list corresponded to 692 cistrons, each containing at least one gene encoding a putative regulator. The interactions predicted with ARACNE were of the type cistron A regulates target cistron B, thus 692 network modules were inferred. Fisher's exact test was then used to identify the network modules in which a significant fraction of genes are associated with a particular functional class or GO term. For the protein classification available from the Sanger Institute (http://www.sanger.ac.uk/Projects/S_coelicolor/scheme.shtml), at a *p*-value threshold of 1.0×10^{-4}, 146 networks were enriched in 33 classes. For the classification using GO terms, at a *p*-value threshold of 1.0×10^{-4}, 115 networks were enriched in 67 GO terms.

4.4.2 Supporting evidence for regulatory networks

4.4.2.1 Network modules containing known edges

The predicted interactions were examined for previously reported interactions identified experimentally. Examples of known interactions include those between pathway specific regulators and their respective biosynthetic clusters. The interactions between *cdaR* (Hojati, Milne et al. 2002; Ryding, Anderson et al. 2002), *actII-ORF4* (Fernandez-Moreno, Caballero et al. 1991) and their corresponding putative clusters (CDA and actinorhodin) were retrieved in our predictions. Regulation of the antibiotic undecylprodigiosin (RED) involves two regulators, *redD* and *redZ* (White and Bibb 1997). RedZ is an activator of *redD* (White and Bibb 1997), which in turn is the activator of the RED biosynthetic pathway. This multilevel regulatory mechanism was retrieved in our results. Interactions were predicted between *redZ* and *redD*, as well as between *redD* and members of the RED cluster. Known interactions involving two-component systems were also predicted, for example, the two-component system AbsA1-AbsA2 acting on the CDA cluster (McKenzie and Nodwell 2007), and VanRS acting on *vanKHAX* (Hong, Hutchings et al. 2004). Other known interactions retrieved in our predictions have been linked to morphology, like that between RamR and *ramCSAB* (O'Connor, Kanellis et al. 2002). The network modules containing these interactions are shown in Figure 4.1.

4.4.2.2 Identification of consensus sequences

Genes which are regulated by a common regulator, often have consensus sequences in their upstream region. MEME was used to examine the upstream region (300 bp) of the first gene in all cistrons belonging to the same network module. A threshold *E*-value (*E*-value$_t$) was determined by identifying consensus sequences in the upstream sequences shuffled randomly. Of the 692 network modules reconstructed with ARACNE, 414 contained a consensus sequence at least in some members with an *E*-

value < *E*-value$_t$. The consensus sequence was present in the upstream region of all the network module members in 84 instances. A list indicating the consensus sequences found in the 414 network modules can be found in Appendix A.

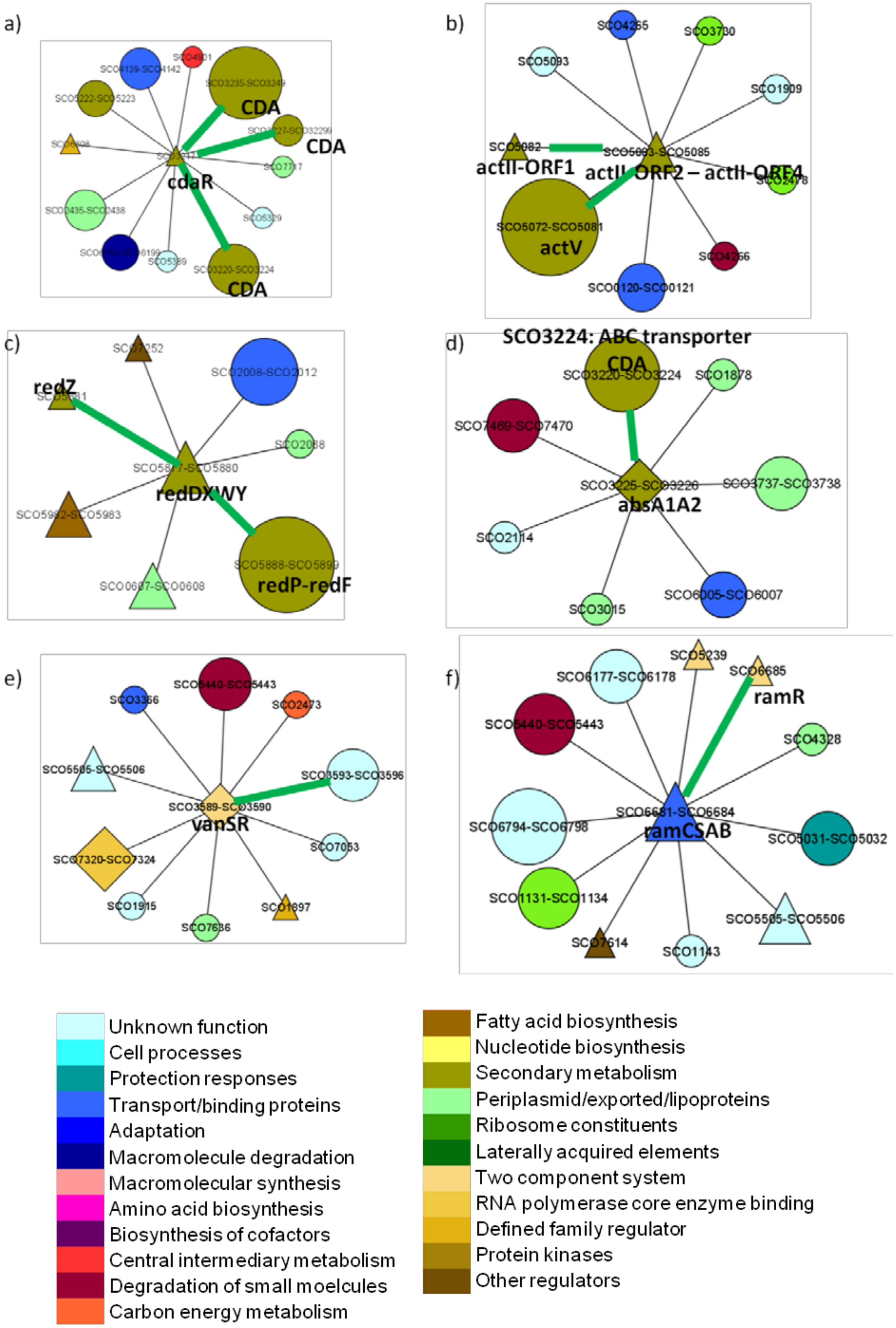

a)
SCO4139-SCO4142
SCO4901
SCO3235-SCO3249
CDA
SCO5222-SCO5223
SCO3227-SCO32299
CDA
SCO6808
SCO3217
cdaR
SCO7717
SCO2435-SCO2438
SCO5329
SCO6198-SCO6199
SCO5389
SCO3220-SCO3224
CDA
b)
SCO4265
SCO3730
SCO5093
SCO1909
SCO5082
SCO5083-SCO5085
actII-ORF1
actII-ORF2 – actII-ORF4
SCO2478
SCO5072-SCO5081
actV
SCO4266
SCO0120-SCO0121
c)
SCO7252
SCO2008-SCO2012
redZ
SCO5881
SCO2088
SCO5877-SCO5880
redDXWY
SCO5982-SCO5983
SCO0607-SCO0608
SCO5888-SCO5899
redP-redF
d)
SCO3224: ABC transporter
CDA
SCO3220-SCO3224
SCO1878
SCO7469-SCO7470
SCO3225-SCO3226
SCO3737-SCO3738
absA1A2
SCO2114
SCO6005-SCO6007
SCO3015
e)
SCO5440-SCO5443
SCO3366
SCO2473
SCO5505-SCO5506
SCO3593-SCO3596
SCO3589-SCO3590
vanSR
SCO7053
SCO7320-SCO7324
SCO1915
SCO7636
SCO1897
f)
SCO5239
SCO6177-SCO6178
SCO6685
ramR
SCO5440-SCO5443
SCO4328
SCO6681-SCO6684
SCO6794-SCO6798
ramCSAB
SCO5031-SCO5032
SCO1131-SCO1134
SCO5505-SCO5506
SCO7614
SCO1143
Unknown function
Cell processes
Protection responses
Transport/binding proteins
Adaptation
Macromolecule degradation
Macromolecular synthesis
Amino acid biosynthesis
Biosynthesis of cofactors
Central intermediary metabolism
Degradation of small moelcules
Carbon energy metabolism
Fatty acid biosynthesis
Nucleotide biosynthesis
Secondary metabolism
Periplasmid/exported/lipoproteins
Ribosome constituents
Laterally acquired elements
Two component system
RNA polymerase core enzyme binding
Defined family regulator
Protein kinases
Other regulators

Figure 4.1. (Previous page) Some network modules with known interactions, among them those between: a) *cdaR* and the CDA cluster; b) *actII-ORF4* and the ACT cluster; c) *redZ* and *redD*, followed by *redD* and the RED cluster ; d) the two-component system AbsA1-AbsA2 and the CDA cluster; e) the two-component system *vanRS* and the VAN cluster; f) morphology related gene *ramR* and the *ramCSAB* operon. Each node represents a cistron and an edge represents an inferred interaction. Node size indicates the size of the cistron. Node shape indicates the number of regulators in the cistron: a circle indicates no regulator, a triangle one regulator, a diamond two or more regulators. Node color indicates functional class. Green edge lines indicate a known interaction.

4.4.2.3 Network modules containing known consensus sequences

Several previous reports on *Streptomyces coelicolor* have identified the upstream consensus binding site of regulatory proteins. The consensus sequences predicted in this study were compared with previously reported binding sites. In several cases an overlap was detected between the consensus sequences. In other cases the consensus sequence was detected in the upstream region of additional network module members. The overlap between consensus sequences determined in this work and previously reported binding sites strengthens the confidence in the predicted network modules and in some cases indicates potential new targets for known regulators.

An ARG box has been reported in the upstream region of the genes *argG* and *argC* (Rodriguez-Garcia, Ludovice et al. 1997). Network module 151, centered on *argRDBJC*, contains eight interactions and it is enriched in the Gene Ontology biological process amino acid metabolic process. A consensus sequence was found in the upstream region of some elements of this network module, including not only *argG* and *argC*, but also *argH* and SCO1086. The previously reported ARG box is shown aligned with the consensus sequence found for network module 151(also shown) in Figure 4.2. In *Streptomyces clavuligerus argGH* is an operon, known to be repressed by *argRDBJC*, an autorepressor. Even though *argG* and *argH* in *Streptomyces coelicolor* are not next to each other, both appear to still be controlled by *argRDBJC*. The upstream region of all three cistrons (*argG*, *argH*, and *argRDBJC*) presents a consensus sequence. SCO1086, annotated as a hypothetical protein, also has the consensus sequence in its upstream region. Prediction of SCO1086 in this network and the presence of the consensus sequence in its upstream region suggest a potential role in arginine metabolism.

An ScbR binding motif has been reported in the intergenic region of *scbA*-*scbR*, which are divergently transcribed, and the upstream region of *kasO* (Takano, Kinoshita et al. 2005). The reconstructed network module 544, centered on *kasO*, contains eight interactions, including interactions with *scbA* and *scbR*. This network module is enriched in the secondary metabolism protein class. It is also enriched in the Gene Ontology molecular functions ligase activity, and cofactor binding. A consensus sequence was found in the upstream region of all nine elements of this network module. The previously reported ScbR binding consensus sequence is shown aligned with the consensus sequence found for the network module (also shown) centered on *kasO* in Figure 4.3. *kasO* is the pathway-specific regulator of the cryptic type I polyketide (TIPK) cluster and its expression is controlled by *scbR* (Takano, Kinoshita et al. 2005). The network module elements SCO6269-SCO6270 and SCO6276-SCO6279 are part of the TIPK cluster. ScbR binds the γ-butyrolactone SCB1 and possibly ScbA. SCB1 triggers production of actinorhodin and undecylprodigiosin (Takano, Nihira et al. 2000). Two cistrons (SCO3227-SCO3229, and SCO3230-SCO3234) from the CDA cluster, also an antibiotic, were predicted as members of this network module; however, the effect of SCB1 on the production of CDA has not been reported. The other network elements, SCO5629 and SCO7192, encode a putative ATP/GTP-binding protein and a putative sigma factor, respectively.

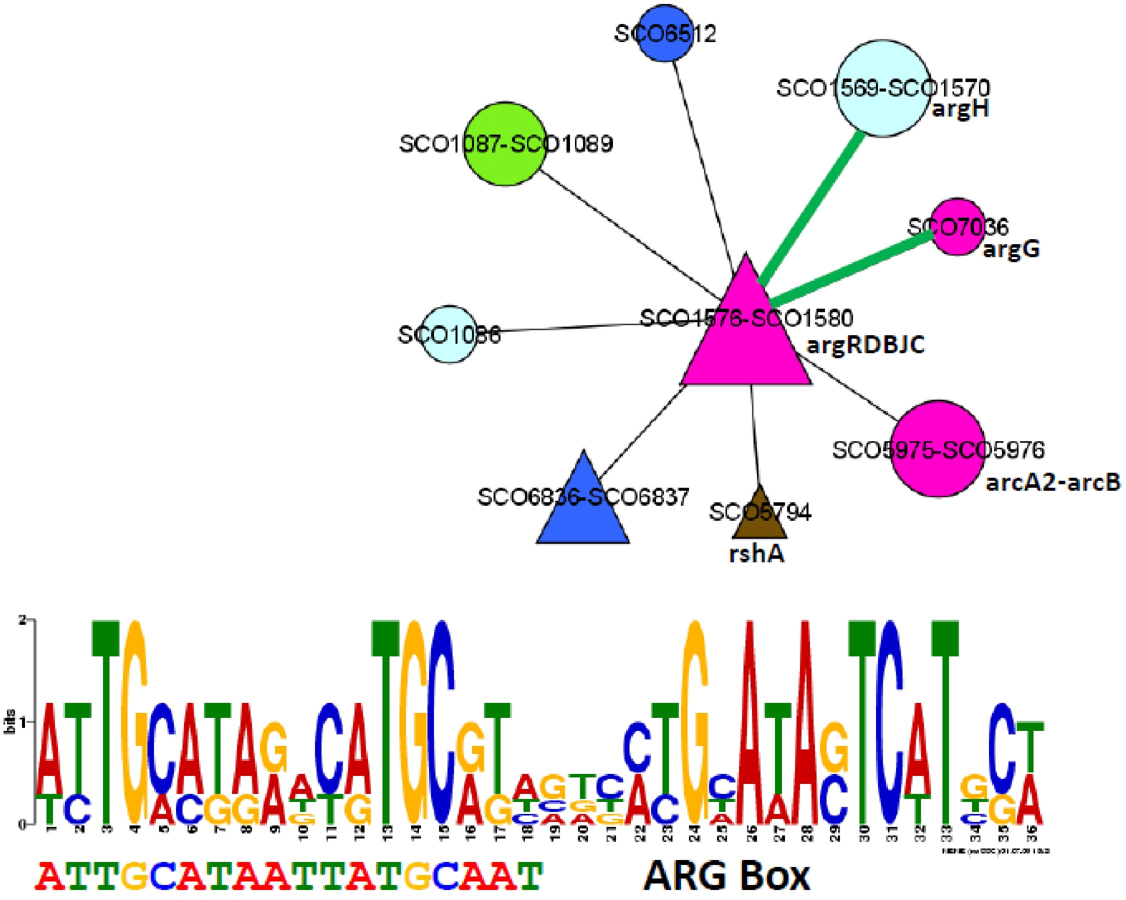


Figure 4.2. Network module 151 centered on *argRDBJC* and consensus sequence identified in this work aligned to the previously reported ARG box. *argGH* is an operon in *Streptomyces clavuligerus*. Node shape and color are as indicated in Figure 4.1.

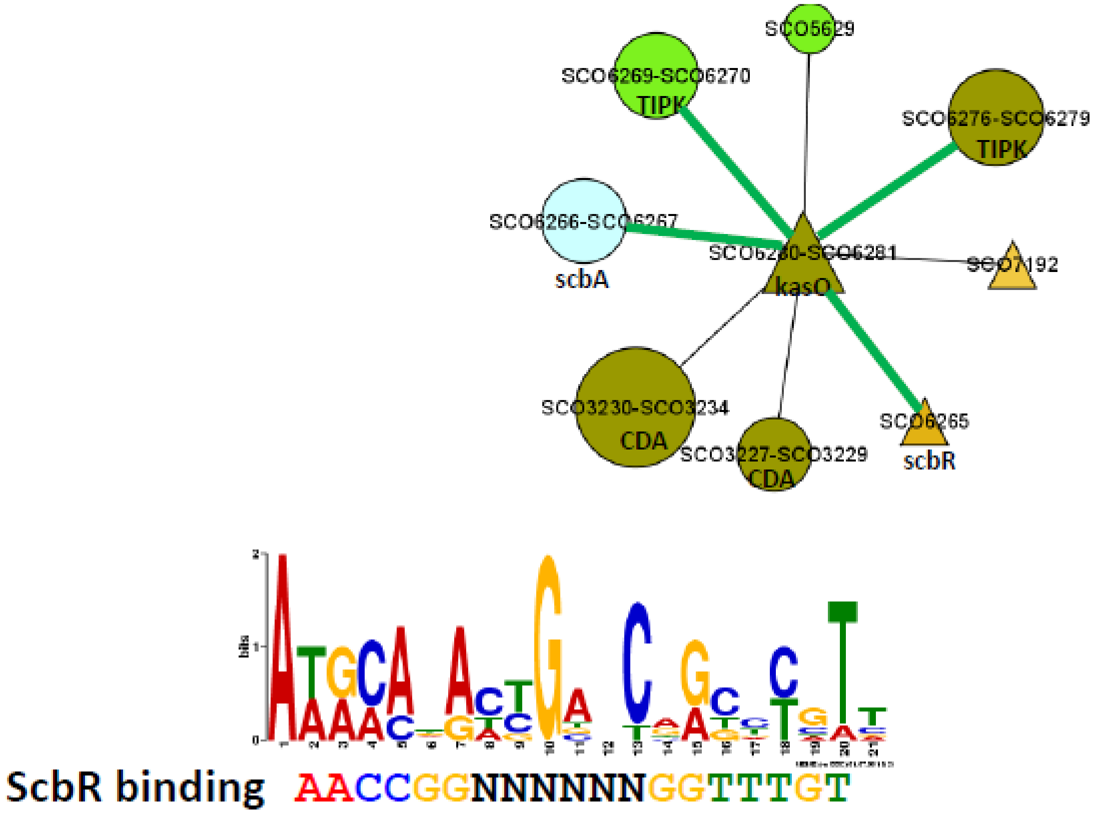


Figure 4.3. Network module 544 centered on *kasO* and consensus sequence identified in this work aligned to the previously reported ScbR binding sequence. This network is related to the γ-butyrolactone system. Node shape and color are as indicated in Figure 4.1.

Two inverted repeats associated with negative control of heat shock genes have been reported, CIRCE (Servant and Mazodier 2001) and HAIR (Grandvalet, de Crecy-Lagard et al. 1999; Servant and Mazodier 2001). CIRCE (Controlling Inverted Repeat of Chaperone Expression), with the consensus sequence TTAGCACTCNNNNNNNNNGAGTGCTAA is present in the upstream region of the *groES*-*groEL1* operon and *groEL2*. The operon *groES*-*groEL1* is part of network module 644, centered on conservon *cvnA13*-*cvnD13*. No consensus sequence passed the *E*-value$_t$ in this network module. *groEL2* is part of network modules 239 and 308. Network module 239 is centered on SCO2950, a DNA-binding protein Hu. This network module did not pass the threshold for functional enrichment, but a consensus sequence was detected in the upstream region of all its elements, including the transcription machinery associated *rpsD* (30S ribosomal protein S4) and *rbpA* (RNA polymerase-associated protein). The motif shows an overlap with the reported CIRCE motif (Figure 4.4). The members of network module 239 could thus potentially represent a module involved in heat shock regulation. The motif from network module 239 is shown in Figure 4.4 aligned with the CIRCE motif and the upstream region of *groEL2*.

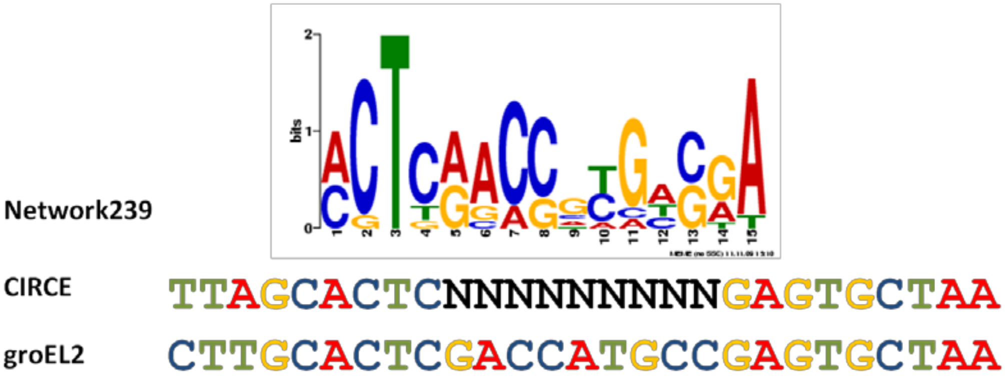


Figure 4.4. Consensus sequence detected in network module 239 aligned to CIRCE (Controlling Inverted Repeat of Chaperone Expression) and the upstream region of *groEL2*.

HAIR (HspR-Associated Inverted Repeat), with the consensus sequence CTTGAGT-N_7-ACTCAAG (Sobczyk, Bellier et al. 2002) is present in the upstream region of *dnaK*, *clpB*, and *lon*. Although close to each other, *dnaK* (SCO3671) and *clpB* (SCO3661) do not form part of the same operon, whereas *lon* (SCO5285) is located distantly. The HAIR element so far has only been found in Gram-positive bacteria with high G+C content. DnaK, ClpB, and GroEL are molecular chaperones. Recently *dnaK*, *clpB*, and *lon* were confirmed as HspR targets using ChIP-chip data (Bucca, Laing et al. 2009). HspR is the repressor of the *dnaK* operon. According to the protein classification scheme, network module 308 centered on the heat shock operon *dnaK*-*grpE*-*dnaJ*-*hspR* is enriched in chaperones, and ribosomal proteins, synthesis modification. This network module is also enriched in the Gene Ontology biological process of protein folding, the cellular component ribosome, and the molecular function structural constituent of ribosome. The network module (Figure 4.5) contains four interactions, including the ribosomal proteins RpsF-Ssb-RpsR*,* the valyl tRNA synthetase ValS*,* the galactose metabolism related proteins GalK-GalE1-GalT*,* and the chaperonin GroEL2. A recent ChIP-chip study (Bucca, Laing et al. 2009) identified new HspR targets, including the ribosomal *rrnD* operon and the $\text{tRNA}^{\text{Gln}}/\text{tRNA}^{\text{Glu}}$ cluster. Based on the newly identified HspR targets, it was suggested that HspR plays a broader role in adapting to stresses than previously hypothesized. This suggestion is strengthened by the prediction of cistrons encoding ribosomal proteins (*rpsF*-*ssb*-*rpsR*) and tRNA synthetase (*valS*) as elements of network 308. A consensus sequence was found in four of the five elements of the network (consensus sequence not present upstream of *valS*). *groEL2* however has been reported to be regulated by HrcA (SCO255) and not by HspR (Bucca, Laing et al. 2009) (Servant and Mazodier 2001).

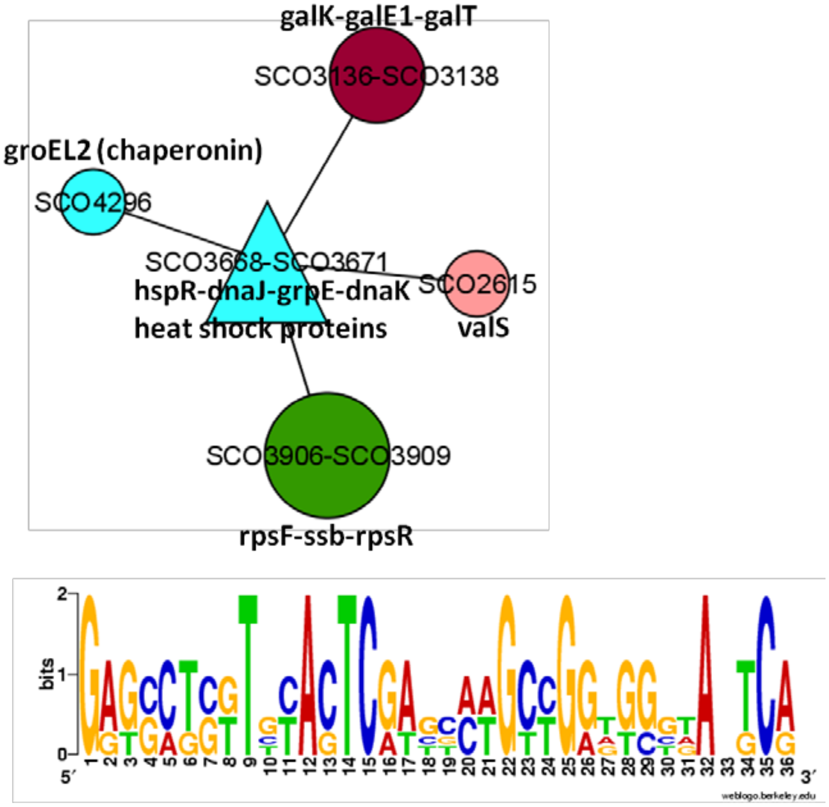


Figure 4.5. Network module 308 centered on *dnaK* and the consensus sequence identified in this work. This network module is related to heat shock. Node shape and color are as indicated in Figure 4.1.

Recently, the *sigU*-dependent promoter sequence TGA[AG]C[AG][N$_{16-17}$]CGTA was identified in the upstream region of 22 likely *sigU* regulon members, many of which encode probable secreted proteins(Gordon, Ottaviano et al. 2008). Furthermore, *sigU*-dependent transcription was tested for nine of the twenty-two elements and confirmed for eight. Network module 240, centered on *sigU* and the gene encoding its regulator, *rsuA* (Gehring, Yoo et al. 2001), contains eighteen interactions. The network module is enriched in the molecular functions ATP-binding, and ATPase activity, coupled to transmembrane movement of substances. Five of the twenty-two proposed *sigU* regulon members are part of network module 240 (SCO0752, SCO0930, SCO2207, SCO2217, and SCO2574), including three of the eight GFP verified genes (SCO0752, SCO0930, and SCO2217). A consensus sequence was identified in seven network module nodes, including the previously proposed *sigU* regulon members (Gordon, Ottaviano et al. 2008) SCO0752, SCO2217, SCO2954, SCO2207, and SCO2575. The consensus sequence was not detected in the previously proposed *sigU* regulon member SCO0930, in which the spacer is seventeen instead of sixteen base pairs. The network module and the identified consensus sequence, aligned with the reported *sigU*-dependent promoter sequence are shown in Figure 4.6.

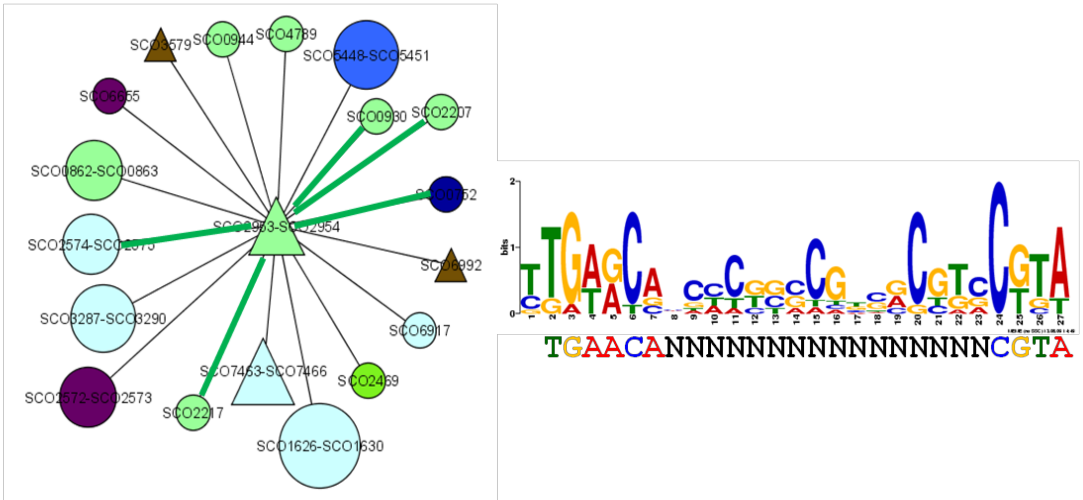


Figure 4.6. Network module 240 centered on *sigU* and the consensus sequence identified in this work, aligned to the previously reported *sigU*-dependent promoter sequence. The *sigU* regulon includes several secreted proteins. Node shape and color are as indicated in Figure 4.1.

The two-component system PhoR-PhoP mediates control of phosphate-regulated genes. PhoP was confirmed to bind the upstream region of *pstS* and the bidirectional promoter region *phoRP*-*phoU*. *pstS* is a phosphate-binding protein precursor and part of the *pstSCAB* operon. *phoU* is a putative phosphate transport system regulator. The six repeats detected in these two promoter regions were used to generate the consensus sequence, denominated PHO box, G(G/T)TCAYYYR(G/C)G (Sola-Landa, Rodriguez-Garcia et al. 2005), where R represents a purine (A or G) and Y a pyrimidine (C or T). Binding of PhoP has also been demonstrated for SCO1565, SCO1968, and SCO7697 (Rodríguez-García, Barreiro et al. 2007). Network modules 370 and 371 were constructed around *phoU* and *phoRP*, respectively. Network module 371 contains only two interactions; one of them is the well known PhoP target *pstSCAB*. The other interaction is with SCO2633, *sodF*, a superoxide dismutase. Network module 371 contains too few elements to be assessed confidently for the presence of a consensus sequence. Network module 370, centered on *phoU*, contains 29 interactions, including the PhoP target *pstSCAB*, and the experimentally validated PhoP targets SCO2878 (Rodríguez-García, Barreiro et al. 2007), and SCO2068, *phoD*, (Apel, Sola-Landa et al. 2007). A consensus sequence was detected in the upstream region of 27 of the 30 network elements, including *phoU*, *pstSCAB*, *phoD*, and SCO2878. The detected motif actually corresponds to the middle part of two direct repeat units (DRu). A PHO box is made of at least two DRus. Both networks and the motif, aligned to a PHO box can be seen in Figure 4.7.

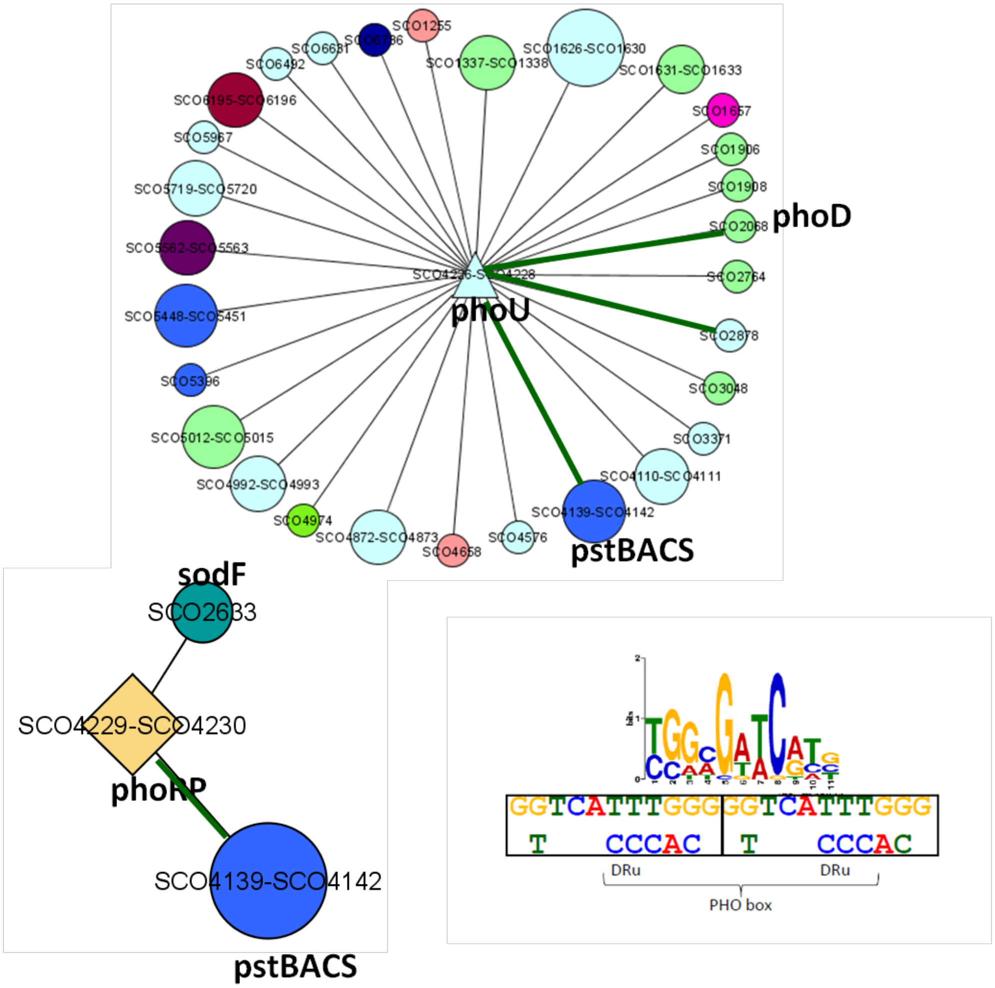


Figure 4.7. Network modules 370 and 371 centered on *phoU* and *phoRP* respectively. These networks are related to the PHO regulon. Node shape and color are as indicated in Figure 4.1.

Studholme et al. (Studholme, Bentley et al. 2004) have used an *in silico* approach to identify eleven sets of functionally coherent genes containing a motif in their upstream regions in the genome of *Streptomyces coelicolor*. The predictions from our study were compared to those eleven sets of genes. The highest overlap was observed between Studholme's matrix 46, which is associated with sigma factors, and our network module 289. Network module 289 is centered on a cistron which contains two genes encoding proteins with regulatory function: SCO3447, a putative transcriptional regulatory protein, and SCO3450, a putative RNA polymerase sigma factor. The overlap between Studholme's matrix 46 and our network 289 includes genes SCO1146, SCO2497, SCO6215, SCO6381, and SCO6994. Genes SCO1146 and SCO6381 code for lipoproteins, whereas the remaining encode hypothetical proteins. Although a consensus sequence is present in 33 of the 39 members of network module 289, including the network hub and the five genes in common with matrix 46, the consensus sequences from the two works do not overlap.

A preliminary consensus sequence present in the promoters of σ^R targets *sigR*, *trxB*, and *hrdD* was used to identify 27 new σ^R target genes (Paget, Molle et al. 2001). The suggested targets were compared to the members of network module 455, centered on *sigR*-*rsrA*. The only gene in common was SCO2161, a conserved hypothetical protein. In an additional study on SigR, SIGffRid (Touzain, Schbath et al. 2008), a tool for searching sigma factor binding sites, was applied to the *Streptomyces coelicolor* and the *Streptomyces lividans* genomes. SIGffRid identifies over-represented patterns in both genomes, by analyzing the promoter regions of orthologous genes. The SigR binding site reported by Paget et al. was searched in the over-represented patterns obtained with SIGffRid. Overlap with the SigR binding sites was detected in the upstream region of 79 genes. The 79 occurrences from that work were compared to the nodes of network module 455. The genes in common, besides *sigR*, included SCO2161, SCO2970, and SCO3404. Because the previously proposed sigma factor binding sites for SigR contain variable spacing, a direct comparison to the consensus sequence from network module 455 was not possible.

4.4.2.4 Overlap of network modules with biological enrichment and consensus sequences

Functionally coherent network modules that also contain a consensus sequence are highly probable to indicate true interactions. Thus, the functionally coherent modules, containing a consensus sequence were identified. A total of 20 network modules contain a consensus sequence in the upstream region of all of its members and present biological enrichment (Table 4.1). Of those network modules, 8 were identified as enriched in both a protein class and a GO term. These network modules will be discussed next.

Table 4.1. Biologically enriched network modules with a consensus sequence in all of its members. MF: Molecular function; BP: Biological process; CC: Cellular component.

Network module	Reg.	Enriched protein class	GO func.	Enriched GO term	Consensus sequence
20	SCO0233	Sec. metabolism	MF	ATPase activity, coupled to transmembrane movement of substances	ACG[AG][TC][GAC][AT]TCA[TC]
42	SCO0422	N/A	MF	Hydrolase activity	G[TC]CGAC[CG][AC]G[CG]T[CG][TG][TA][CT]
45	SCO0453	Transport/binding proteins	MF	Hydrolase activity, hydrolyzing O-glycosyl compounds Transporter activity	[CG]ACCG[CAG]C[AC][AT][GT]CTC[GCT][CA]C[GA][TC][GC][ACG][AC]
			BP	Transport	
49	SCO0487	Sigma factor	MF	DNA binding Sigma factor activity	[GC][GA][TCG]G[GA]T[CG]A[CG][CG][GT]
			BP	Transcription initiation	
56	SCO0588	Anaerobic respiration Electron transport Fatty acid and phosphatidic acid biosynthesis	MF	Electron transporter activity NADH dehydrogenase (ubiquinone) activity Nitrate reductase activity	[TA][CG][GC][TA]CG[TA][CGT][CG][ACG]CG[AG][CG][GT][TA]C[GT][TC]CG
			BP	Mitochondrial electron transport, NADH to ubiquinone	
			CC	Nitrate reductase complex	
110	SCO1099	N/A	MF	NADH dehydrogenase (ubiquinone) activity	[GA][TC]G[AG][TC][CG][AT][CA]G[AC][AT][CG]T[ACGTA]C
118	SCO1177	Not classified	N/A	N/A	CGGC[AG]T[CG]TCGT
219	SCO2565	N/A	MF	DNA-directed DNA polymerase activity	[CG]A[CG]G[AG]CGA[CG]G[AT]C[CG]T
			BP	DNA replication	
246	SCO3014	N/A	MF	Kinase activity Transferase activity, transferring phosphorus-containing groups	GTTCA[TCA][CG][TG][CT]
			BP	Phosphorylation	
358	SCO4124	Plasmid-related functions	N/A	N/A	G[TG]CG[AC]C[CG][AG]G[GC][GAC]C[GC][AT]G

Network module	Reg.	Enriched protein class	GO func.	Enriched GO term	Consensus sequence
397	SCO4477	N/A	BP	Aromatic compound metabolic process	[GC]A[GA]G[TC]C[GA]T
431	SCO4920	Sigma factor	MF	Sigma factor activity	T[GT][TA]TCG[AC]
			BP	Transcription initiation	
494	SCO5692	Plasmid-related functions	N/A	N/A	[TCG]T[CG][CG][AG][CT][CG][AT][CG]G[AG][CG][CG][TCG][GCA]G[GC][CG][GC][AG][TCG][CGT][GTC]G[GC][AC][GC]G[TA]C[CG][TG][CG][GCT][TAG][CA][GC][GA][TC]G[AT]
544	SCO6280	Sec. metabolism	MF	Ligase activity Cofactor binding	A[TA][GA][CA][AC][GTA][AG][CAT][TC]GA[CAGTA]C[AG][GA][CGT][CT][CT][GT]T[TC]
608	SCO7086	N/A	BP	Aromatic compound metabolic process	[TC]CGTC[GC][GT][CTG][GC][AC]
636	SCO7325	Adaptations, atypical conditions	MF	Structural molecule activity	[TC][GC][AG]T[GC][TA][AT]CA
646	SCO7497	Festorage	N/A	N/A	C[CG][TG]TC[CAT][TGA]C[GA][TG]C
684	SCO7759	N/A	MF	Transporter activity	CGA[CG][CT][GTC]G[CT]AC[GT][AT][CAG][CG][TG]
685	SCO7765	Sec. metabolism	N/A	N/A	GA[GC][AGC][ACT][CGA]GTC[AGCTC][TA][CT][ACT]T[CT]
691	SCO7808	Cobalamin	MF	Methyltransferase activity	[CA]CG[TC][TC]G[ACG]T[CG][AT]C
			BP	Biosynthetic process	

The most noteworthy case is that of network module 56, centered on SCO0588. SCO0588 is a putative sensor kinase within conservon *cvnABCD11*. Conservons are conserved operons containing four (or in some cases five) genes. Thirteen conservons were identified in the *Streptomyces coelicolor* genome (Bentley, Chater et al. 2002). Network module 56 presents enrichment in anaerobic respiration, electron transport, and fatty acid and phosphatidic acid biosynthesis, according to the protein classification scheme. The network module also shows enrichment in the Gene Ontology molecular functions NADH dehydrogenase (ubiquinone) activity, and nitrate reductase activity, and the cellular component nitrate reductase complex. With 53 edges this is the network module with the highest number of edges. The network module and the consensus sequence, which is present in all the network module members, are shown in Figure 4.8.

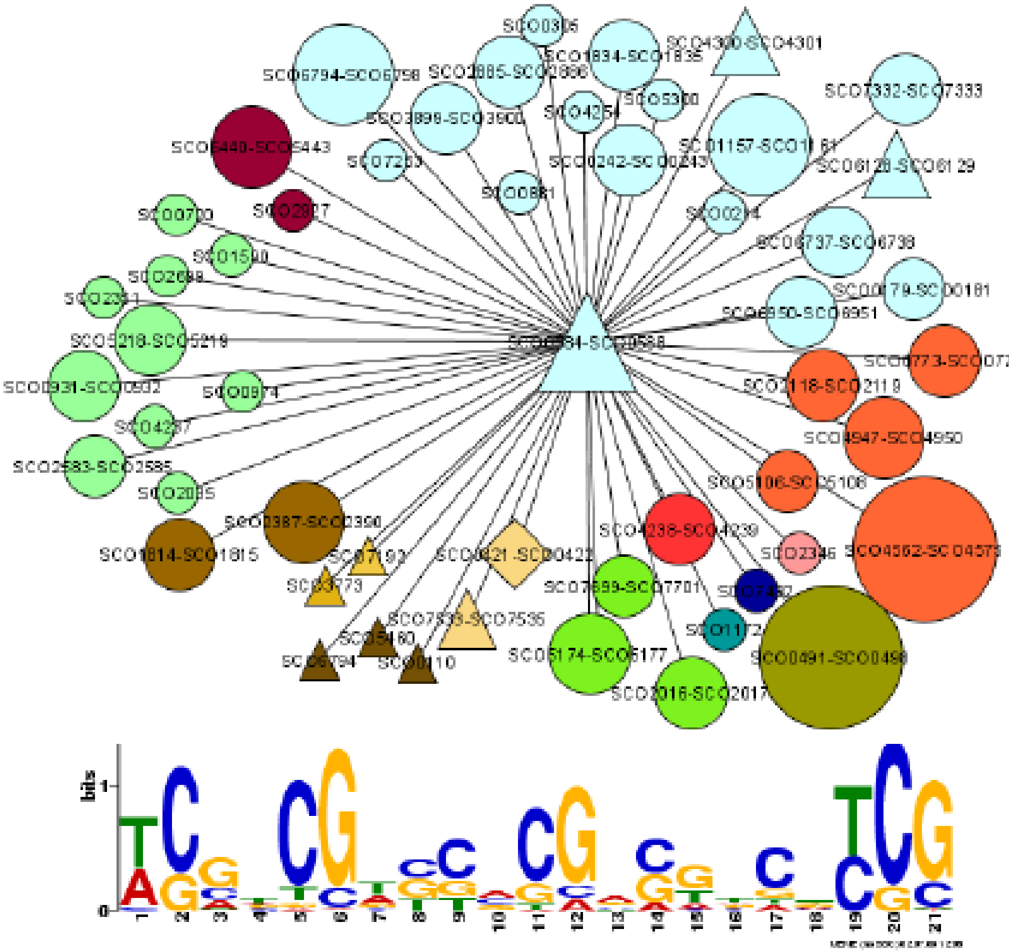


Figure 4.8. Network module 56 centered on SCO0588 and the consensus sequence detected in all the network module members. This is the most highly connected network, with 53 interactions. Node shape and color are as indicated in Figure 4.1.

Following in number of edges are network module 49 (24 elements) and network module 431 (20 elements). Both network modules are centered in family transcriptional regulators and present enrichment in the same protein class and Gene Ontology terms. The protein class enriched is sigma factor; the Gene Ontology terms enriched are sigma factor activity, and transcription initiation.

The center of network module 49, SCO0487 is a putative MarR family transcriptional regulator adjacent to the cluster for the chelator coelichelin (SCO0489-SCO0499). Most transcriptional regulators of the MarR (multiple antibiotic resistance regulator) family are autoregulators. MarR regulates genes in response to environmental stress, virulence factors, and aromatic catabolic pathways (Wilkinson and Grove 2006). One third of the 24 cistrons of network module 49 contain regulatory genes, including extracytoplasmic function (ECF) sigma factors (SCO0159, SCO0866, and SCO1263), a sigma factor (SCO6239), a two-component system (SCO3351-SCO3352), a DNA binding protein (SCO3352), a regulatory protein (SCO6237), and the MarR family regulator (SCO0487) itself. ECF sigma factors are in general cotranscribed with a negative regulator, often acting as an anti-sigma factor. After detecting a signal from the environment the sigma factor is released and can bind to RNA polymerase and activate transcription. This mechanism is analogous to that of two-component systems, and both are mechanisms for coordinating cellular responses to external signals (Helmann 2002). The presence of a high number of ECF sigma factors, a two-component system, and a family transcriptional regulator, all involved in response to environmental changes, plus the presence of proteins classified as involved in transport/binding (SCO0448-SCO0449, SCO3330-SCO3331, SCO3703, SCO6690, and SCO7503-SCO7505), indicate the possibility that network module 49 is involved in responding to some as yet undetermined environmental change. This network module is shown in Figure 4.9.

The center of network module 431, SCO4920 is a putative deoR family transcriptional regulator. This defined-family regulators have been reported to regulate genes involved in carbohydrate metabolism (Sakakibara and Saha 2008). More than one third of the cistrons in network module 431 include regulatory genes. The sigma factors present in this network module include the morphology related *bldN*, the putative RNA polymerase sigma factor SCO4409, and the ECF sigma factor SCO4769. This network module includes conservon *cvnABCDE9*, a probable membrane-associated complex which may connect to the *bld* cascade (Komatsu, Takano et al. 2006). Network module 431 is shown in Figure 4.10.

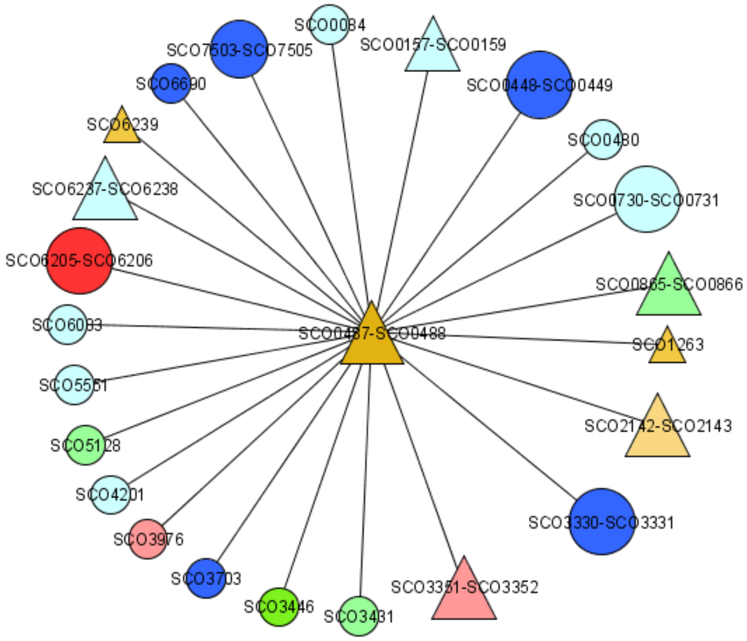


Figure 4.9. Network module 49, centered on SCO0487. Eight of the twenty-four network module members encode regulators (triangles). Node shape and color are as indicated in Figure 4.1.

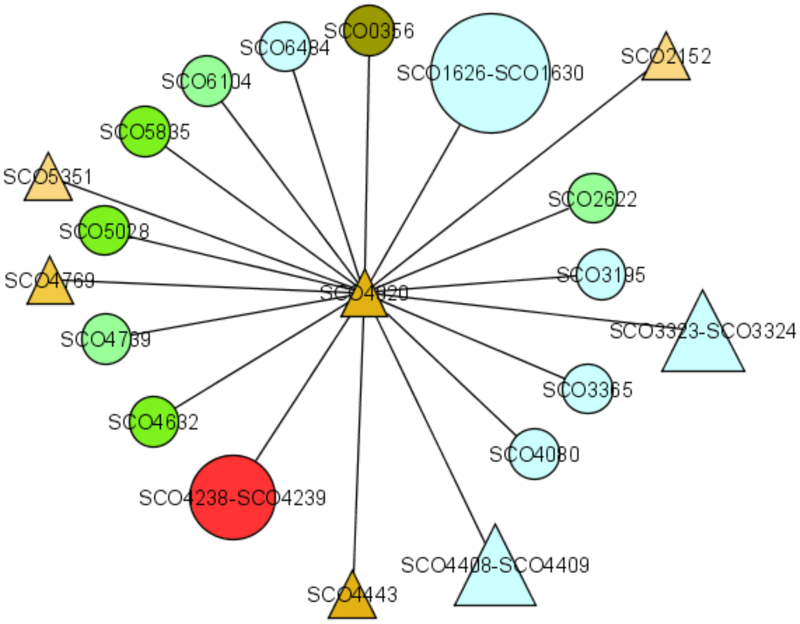


Figure 4.10. Network module 431 centered on SCO4920. This network module contains a high number of regulators (triangles). Node shape and color are as indicated in Figure 4.1.

Network modules 20 and 544 both present a consensus sequence in the upstream region of all their elements and appear enriched in the protein class secondary metabolism. Both network modules also present enrichment in at least a Gene Ontology term. Network module 544 was discussed previously (Section 4.4.2.3). Network module 20 is centered on the putative DNA-binding protein SCO0233. This network module presents enrichment in the Gene Ontology molecular function ATPase activity. The cistron SCO7681-SCO7691 corresponds to genes that are part of the coelibactin cluster. Two recent studies have reported that both *zur* (SCO2808) and *absC* (SCO5406) control expression of the coelibactin cluster; overlapping sites were found for direct binding of Zur and AbsC in the intergenic region between SCO7681 and SCO7682 (Hesketh, Kock et al. 2009; Kallifidas, Pascoe et al.). In both studies SCO0472-SCO0475, also part of network module 20, was also reported as affected by mutations in *absC* and *zur*. A Zur binding site was reported in the region between SCO0475 and SCO0476. Network module 20 is shown in Figure 4.11.

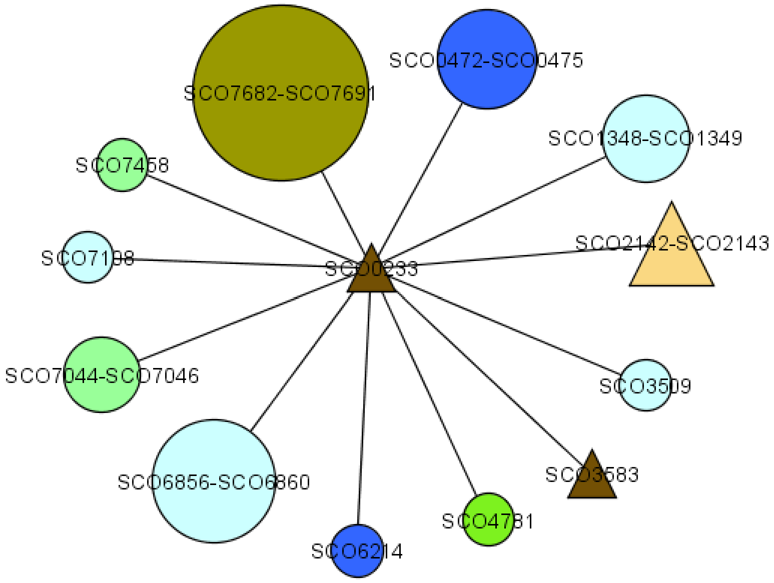


Figure 4.11. Network module 20 centered on SCO0233. Node shape and color are as indicated in Figure 4.1.

Other smaller networks including a consensus sequence in all their members and enrichment in a protein class and a Gene Ontology term include network modules 636 (9 elements), 691 (9 elements), and 45 (8 elements). Network module 636 is centered on SCO7325, an anti-sigma factor antagonist. The network is enriched in the protein class adaptations, and the Gene Ontology molecular function structural molecule activity. The biggest cistron in this network, SCO6499-SCO6508, encodes the putative gas vesicle synthesis proteins GvpOAFGYZJLSK. Gas vesicles function as flotation devices; however the involvement of gas vesicle proteins in processes other than flotation has been suggested for actinomycetes (van Keulen, Hopwood et al. 2005; Walsby and Dunton 2006). Network module 636 contains regulatory genes (SCO4434, SCO6003, SCO7325, and SCO7446); secreted or membrane proteins (SCO0297, SCO3802, and SCO6002); and hypothetical proteins (SCO0775, and SCO7753). Unfortunately none of this can confirm the function of *gvp* genes, nor suggests a new function for them.

Network module 691 is centered on SCO7806, a putative DNA-binding protein. The network module is enriched in the protein class cobalamin, the Gene Ontology molecular function methyltransferase activity, and the Gene Ontology biological process biosynthetic process. The cistron SCO1855-SCO1859 contains two methyltransferases (SCO1855 and SCO1856) and an aminotransferase (SCO1859).

The last network module in which a consensus sequence was detected in all its members and that presents enrichment in both a protein class and a Gene Ontology term is network module 45, centered on SCO0456, a LacI family transcriptional regulator. This network module is enriched in transport using both the protein classification and Gene Ontology terms. The transport related proteins include SCO0454, SCO0455, and the sugar transport proteins SCO0538, SCO0539, and SCO0540.

4.4.3 Bidirectional promoters

A bidirectional gene pair is formed by two adjacent genes located on opposite strands of DNA, that is, the genes are arranged head-to-head (Wang, Wan et al. 2009). These divergently transcribed genes are common in the human genome (Adachi and Lieber 2002) and are well conserved among prokaryotes (Korbel, Jensen et al. 2004). Functional relationship has been reported for some bidirectional pairs.

Bidirectional promoters present coregulation opportunities. Binding sites for regulatory proteins in the bidirectional promoter region may regulate both genes; with activation/repression of both genes, or activation of one gene and repression of the other (Beck and Warren 1988).

The genome of *Streptomyces coelicolor* contains 1429 pairs of divergently transcribed genes, however transcript start has been determined for only a couple of cases. Using protein start positions we identified 17 cases in which the coding region of divergently transcribed genes overlaps. These cases will be referred to as DO, for Divergently transcribed Overlapping coding regions. We also identified 63 cases in which the protein start of genes in opposite strands is between one and thirty-five nucleotides. These cases will be referred to as DNO, for Divergently transcribed Non Overlapping coding regions. Whereas none of the DO pairs were identified in the same network, two of the DNO pairs were (Table 4.2). The first case is that of the pair formed by SCO3353 (hypothetical protein) and SCO3354 (hypothetical proline-rich protein), both part of network module 344, centered on SCO4008 (putative tetR family regulatory protein). Network module 344 is enriched on the GO term regulation of transcription, DNA dependent. No consensus sequences passed the *E*-value$_t$ for this network module. The second case is that of the pair formed by the cistrons SCO4672-SCO4673 and SCO4674. SCO4672 encodes a putative secreted protein, SCO4673 a putative deoR-family transcriptional regulator, and SCO4673 a hypothetical protein. SCO4672-SCO4673 is the center of network module 411. Network module 411 does not present functional enrichment nor consensus sequences.

Some *Streptomyces coelicolor* bidirectional pairs previously studied, like *scbA*-*scbR* and *phoRP*-*phoU*, which are known to interact, present bigger distances between their protein start positions (118 and 215 bp respectively). Furthermore, the distance between the protein start positions for the *E. coli* bidirectional pair *maltT*-*malPQ* is 611 bp (Beck and Warren 1988). The analysis of bidirectional pairs was thus extended to include distances of up to 300 bp. Of the 1429 divergently transcribed gene pairs in the *Streptomyces coelicolor* genome 1153 of them have a distance of up to 300 bp between their corresponding protein starts.

Of the 1153 pairs thirteen appear in the same network (Table 4.2) with neither of them being the network center. In five of these cases a motif is present in both elements in the pair plus the cistron in which the network is centered (SCO6265 and SCO6266; SCO0170 and SCO0171; SCO6693 and SCO6694; SCO3229 and SCO3230; and SCO5286 and SCO5287). Network module 544, centered on *kasO* and previously discussed (Section 4.4.2.3, contains the well known case SCO6265-SCO6266 (ScbA-ScbR) but in addition contains the bidirectional pair SCO3229-SCO3230, located within the CDA biosynthetic cluster. Network module 544 is enriched in the secondary metabolism protein class and a consensus sequence was found in the upstream region of all the network members.

Of the 1153 pairs 18 appear in the same network (Table 4.2) with one of them being the network center. Four of those pairs (SCO0193 and SCO0194; SCO5082 and SCO5083; SCO6368 and SCO6369; and SCO7614 and SCO7615) appear in two separate networks, since both pair elements correspond to genes encoding regulators and each is the center of a different network. A consensus sequence was detected in nine of these bidirectional pairs (SCO0116 and SCO0117; SCO3224 and SCO3225; SCO4187 and SCO4188; SCO6357 and SCO6358; SCO7489 and SCO7490; SCO0193 and SCO0194; SCO5082 and SCO5083; SCO7614 and sCO7615; and SCO6265 and SCO6266). Three of the bidirectional pairs are related to secondary metabolism: SCO3225 and SCO3224 (CDA cluster), SCO5082 and SCO5283 (ACT cluster) and SCO6265 and SCO6266 (ScbA-ScbR). Correlation between both elements in the 18 pairs is strong, as evidenced by their MI values, with the exception of pair SCO0193-SCO0194 (MI < 0.10). Furthermore, six pairs (SCO3224-SCO3225; SCO7489-SCO7490; SCO6357-SCO6358; SCO4677-SCO4678; SCO0064-SCO0065; and SCO5082-SCO5083) have an MI > 0.163; which is the value of the ScbA-ScbR pair.

Table 4.2. Summary of bidirectional pairs with a distance less than 300 bp and that are part of the same network module.

First gene	Second gene	Distance (bp)	Network	Type	MI	Motif
SCO0064	SCO0065	292	3	center-edge	0.235	N
SCO0116	SCO0117	117	6	center-edge	0.154	Y
SCO0170	SCO0171	210	14	edge-edge	NA	Y
SCO0193	SCO0194	161	15 16	center-edge	0.072	Y Y
SCO0916	SCO0917	251	80	edge-edge	NA	N
SCO1751	SCO1752	262	201	edge-edge	NA	N
SCO1802	SCO1803	218	167	center-edge	0.121	N
SCO3005	SCO3006	282	365	edge-edge	NA	N
SCO3224	SCO3225	137	269	center-edge	0.164	Y
SCO3229	SCO3230	269	544	edge-edge	NA	Y
SCO3353	SCO3354	11	344	edge-edge	NA	N
SCO3505	SCO3506	161	289	edge-edge	NA	N
SCO3834	SCO3835	159	329	center-edge	0.132	N
SCO4187	SCO4188	154	364	center-edge	0.102	Y
SCO4673	SCO4674	12	411	center-edge	0.112	N
SCO4677	SCO4678	113	412	center-edge	0.233	N
SCO4944	SCO4945	169	433	center-edge	0.126	N
SCO5082	SCO5083	111	445 446	center-edge	0.358	Y Y
SCO5122	SCO5123	187	245	edge-edge	NA	N
SCO5286	SCO5287	249	637	edge-edge	NA	Y
SCO5662	SCO5663	57	48	edge-edge	NA	N
SCO5783	SCO5784	115	503	center-edge	0.130	N
SCO6265	SCO6266	118	542 544	center-edge edge-edge	0.163 NA	Y Y
SCO6357	SCO6358	118	553	center-edge	0.203	Y
SCO6368	SCO6369	159	555 556	center-edge	0.144	N N
SCO6693	SCO6694	176	30	edge-edge	NA	Y
SCO7042	SCO7043	85	518	edge-edge	NA	N
SCO7489	SCO7490	126	645	center-edge	0.186	Y
SCO7614	SCO7615	99	662 663	center-edge	0.129	Y Y
SCO7753	SCO7754	92	683	center-edge	0.123	N

4.5 Discussion

Transcriptional control of gene expression is a fundamental process in prokaryotes. However, the gene regulatory circuits are not well understood. Temporal gene expression profiles procured under diverse genetics and environmental perturbations can provide valuable information for deciphering the mechanisms of interplay between the regulatory genes and their ensuing effects. In this work regulatory network modules inferred from transcriptome data at the level of cistrons were analyzed for their biological meaning. Some of the interactions in the network modules correspond to experimentally known interactions. In addition, the network modules were analyzed for functional enrichment and the presence of consensus sequences. Some of the consensus sequences overlap previously described binding sequences and motifs.

The relevance of secondary metabolism in this species was reflected in the high number (25) of network modules enriched in the secondary metabolism protein class. Additionally, 16 networks were found to be enriched in the polyketide synthase functional class. Other protein classes that are enriched in multiple networks include amino acids and amines (13 networks), anaerobic respiration (12 networks), electron transport (9 networks), and adaptations, atypical conditions (8 networks).

The Gene Ontology term that appeared as enriched in the most number of networks was NADH dehydrogenase (ubiquinone) activity, enriched in 13 networks. Nitrate reductase activity, and nitrate reductase complex followed, each enriched in nine networks.

A total of twenty networks presented functional enrichment and the presence of a consensus sequence in all of its members. These twenty networks offer the highest probability of representing true interactions.

4.6 Concluding remarks

Even though the inferred networks were based only on transcriptome data it was encouraging that known protein-DNA interactions were obtained in several networks. In the future, not only could EMSA be used to confirm some of the predicted interactions, but other techniques like ChIP-chip and even ChIP-Seq could also be used. Currently however, genome-scale experimental data for protein-DNA interactions in *Streptomyces coelicolor* is almost non-existing. This approach demonstrates the value of data mining once a large data set is available and provides information that can help in guiding experiments, by suggesting targets for a regulator that has not been studied so far or additional targets for previously studied regulators.

Chapter 5
Genome Sequencing of the Curdlan Producer *Agrobacterium* sp. ATCC 31749

5.1 Summary

Agrobacterium sp. ATCC 31749 produces the exopolysaccharide curdlan, a gelling agent with uses in the food industry. Genetic information on this organism, however, is limited. Sequence information for genes involved in curdlan production is almost nonexistent. In this work, we sequenced the genome of *Agrobacterium* sp. ATCC 31749, resulting in a draft genome. The draft genome was compared to the genome of close relatives, both at a global level and at the gene level for those genes known to have a role in curdlan production. The genetic information obtained in this project allowed our collaborators at the Georgia Institute of Technology to create a custom oligonucleotide microarray which is being used to identify target genes for metabolic engineering and to characterize the transcriptome of the resulting mutants.

5.2 Introduction

5.2.1 The Rhizobiaceae family

The genera *Agrobacterium* belongs to the Rhizobiaceae family, which also includes the genera *Rhizobium*, *Allorhizobium*, *Ensifer* and *Sinorhizobium*. There has been controversy as to whether the genera *Agrobacterium*, *Allorhizobium* and *Rhizobium* represent distinct phenotypic entities or not. Young et al. (Young, Kuykendall et al. 2001) proposed to amalgamate these three genera into the unique genus *Rhizobium*, renaming *Agrobacterium* to *Rhizobium*. This proposal, however, has generated some controversy (see for example (Farrand, Van Berkum et al. 2003; Young, Kuykendall et al. 2003)).

Rhizobia can grow in soil as free-living organisms, but can also be found as symbionts inside root nodule cells of legumes (Gage 2004). Inside legume-root nodules, Rhizobia reduce dinitrogen to ammonium. The ammonium is secreted to the plant, which in exchange provides Rhizobia with carbon and energy sources, mostly as malate and succinate (Prell and Poole 2006).

Genome organization among Rhizobia is diverse, including one circular chromosome, two circular chromosomes, one circular chromosome and one linear chromosome, plus plasmids (Sobral, Honeycutt et al. 1991; Allardet-Servent, Michaux-Charachon et al. 1993; Jumas-Bilak, Michaux-Charachon et al. 1998). As the presence of essential genes has not been analyzed in some replicons, in some cases there is uncertainty if the elements correspond to macro plasmids or chromosomes. Plasmids do not contain essential genes. Genes required for symbiosis are usually located in large plasmids, although they can be chromosomally located too, in regions known as symbiotic islands (Ding and Hynes 2009).

5.2.1.1 Agrobacterium sp. ATCC 31749

Agrobacterium sp. ATCC 31749 is a Gram-negative, aerobic, rod-shaped, non-phytopathogenic bacterium that produces a gelable exopolysaccharide known as curdlan. It was originally deposited as *Alcaligenes faecalis* subsp. myxogenes. *Agrobacterium* sp. ATCC 31749 has been engineered to produce the medically-relevant Gal-α1,3-Lac by Ruffing et al. (Ruffing and Chen 2010).

5.2.1.2 Agrobacterium tumefaciens **C58**

Agrobacterium tumefaciens C58 is a plant pathogen that causes tumor formation in most dicotyledonous and some monocotyledonous species (Pitzschke and Hirt 2010). Its genome was sequenced in 2001 by two groups, one at the University of Washington (Wood, Setubal et al. 2001), and the other at Cereon (Goodner, Hinkle et al. 2001). Its total genome is 5.67 Mb and consists of a circular chromosome (2.8 Mb), a linear chromosome (2.1 Mb), the plasmid At (543 Kb), and the plasmid Ti (214 Kb). The genes required for tumor formation (*vir* genes) are encoded in the Tumor-inducing (Ti) plasmid. The transfer DNA (T-DNA) is a discrete set of genes which are imported into plant cells and integrated into their chromosomal DNA. The T-DNA region is also located in the Ti plasmid. The genes within the T-DNA can be replaced by any other DNA sequence, making *Agrobacterium tumefaciens* C58 a natural genetic engineer.

5.2.2 Curdlan production by *Agrobacterium* sp. ATCC 31749

Curdlan is a linear (1→3)-β-glucan which produces reversible gels when its aqueous solutions are heated to 55°C and irreversible gels when heated to temperatures above 80°C (McIntosh, Stone et al. 2005). Curdlan is used as a food additive in dairy products, and as a stabilizer and thickener in processed foods. In addition to its uses in the food industry curdlan has potential uses in medicine. Curdlan sulfate has shown anti-HIV activity (Jagodzinski, Wiaderkiewicz et al. 1994).

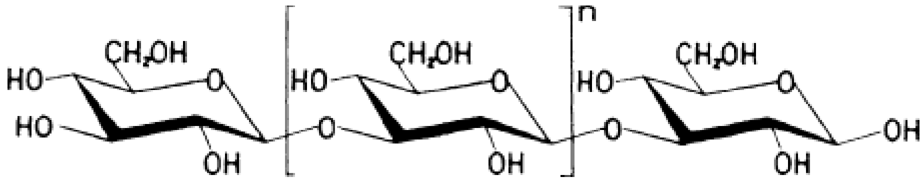


Figure 5.1. Structure of curdlan, from (McIntosh, Stone et al. 2005)

Curdlan is produced by strains of *Agrobacterium ATCC 31749*, *Agrobacterium radiobacter*, *Agrobacterium rhizogenes, Rhizobium trifolii* J60, and *Rhizobium* sp. TISTR 64B (Nakanishi, Kimura et al. 1976; Ghai 1981; McIntosh, Stone et al. 2005). Curdlan synthesis in cultures is triggered by nitrogen depletion (Phillips and Lawford 1983; Phillips, Pik et al. 1983), mimicking the production of secondary metabolites (Saudagar and Singhal 2004).

Genes for the production of curdlan in *Agrobacterium* sp. ATCC 31749 were first cloned by Stasinopoulos et al. (Stasinopoulos, Fisher et al. 1999). They found four genes essential for curdlan production: *crdA*, *crdS*, *crdC*, and *crdR.* An additional gene, a phosphotidyl serine synthase, pss_{AG}, is required for maximal yields (Karnezis, Fisher et al. 2002). Nitrogen metabolism genes, *ntrBC*, have an effect on curdlan production.

The operon *crdASC* occupies a 4,948 bp region, and the flanking genes are transcribed in the opposite direction (Karnezis, Epa et al. 2003). The only sequence publicly available however is that of *crdS*. This made any attempts to perform metabolic engineering on this strain very limited.

CrdS is an inner membrane protein with seven transmembrane helices, one non-membrane-spanning amphipatic helix and an N_{out}-C_{in} disposition (Karnezis, Epa et al. 2003). *crdS* is flanked by *crdA* and *crdC*. These two genes present no homology to reported proteins sequences. Their role in the production of curdlan is still unknown, but it was hypothesized that CrdA and CrdC might form a multimeric complex with CrdS that might help in the transport of the polymer (Karnezis, Epa et al. 2003). The location of the fourth gene essential for curdlan production, the probable transcriptional activator CrdR, is not available.

5.2.3 Curdlan gene orthologues in other *Agrobacterium*

The genome of the plant pathogen *Agrobacterium tumefaciens* C58 has a *crdS* orthologue (Atu3056). The genes flanking *crdS* in *Agrobacterium tumefaciens* C58 are simply annotated as hypothetical proteins. Preliminary information from the *Agrobacterium radiobacter* K84 genome sequencing project obtained through the Agrobacterium.org website (http://depts.washington.edu/agro/) indicates that *Agrobacterium radiobacter* K84 does have a orthologue to *crdS* (Arad7561) but no flanking genes that could correspond to *crdA* and *crdC* (Figure 5.2). *Agrobacterium radiobacter* K84 has been reported to produce curdlan, whereas *Agrobacterium tumefaciens* C58 has not.

The genes *rcdA* and *rcdB*, present in both *Agrobacterium tumefaciens* C58 (Atu5090-Atu5091) and *Agrobacterium radiobacter* K84 (Arad7561-Arad7562), are also annotated as curdlan related. RcdA is annotated as a curdlan synthase and RcdB as a curdlan synthesis protein. In both organisms, *rcdAB* are transcribed in the same direction (Figure 5.3). Note from Figure 5.2 that RcdA from *Agrobacterium radiobacter* K84 (Arad7561) was given as an orthologue to Atu3056, the *crdS* orthologue from *Agrobacterium tumefaciens* C58.

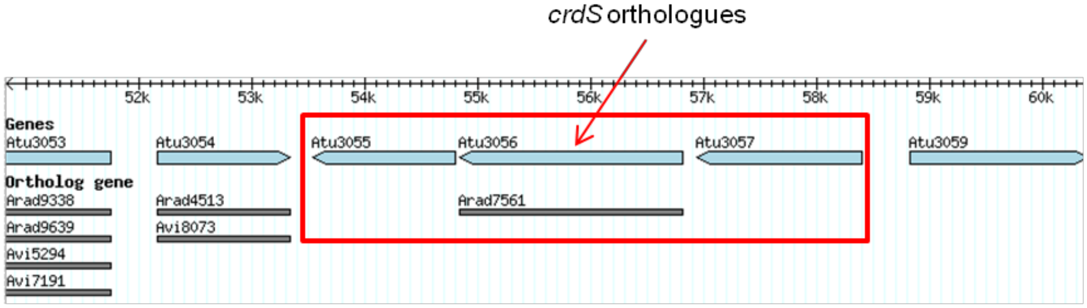


Figure 5.2. *crdS* orthologue (Atu3056) and flanking genes in *Agrobacterium tumefaciens* C58. Atu3056 has an orthologue in *Agrobacterium radiobacter K84 (Arad7561)*. Modified image from (http://depts.washington.edu/agro/).

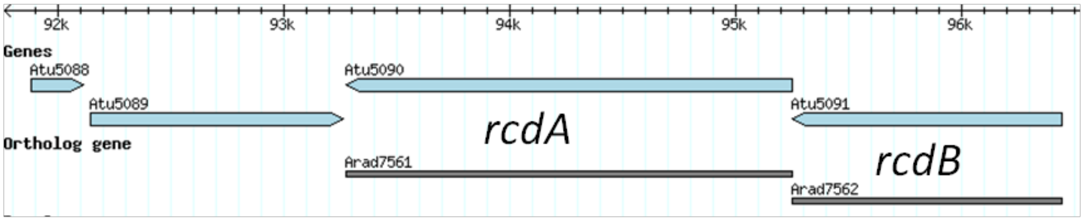


Figure 5.3. *rcdA* (Curdlan Synthase) and *rcdB* (Curdlan synthesis protein) in *Agrobacterium tumefaciens* C58 and *Agrobacterium radiobacter* K84. Modified image from (http://depts.washington.edu/agro/).

5.3 Materials and Methods

5.3.1 gDNA preparation and sequencing

Agrobacterium sp. ATCC 31749 cultures for genomic DNA (gDNA) extraction were carried out by Dr. Anne Ruffing at the Georgia Institute of Technology. Briefly, *Agrobacterium* sp. ATCC 31749 was cultivated overnight in 4 mL of LB media at 30°C with agitation at 250 rpm. gDNA extraction was done using the GenElute Bacterial Genomic DNA Kit (Sigma-Aldrich, St. Louis, MO). Library preparation and sequencing were performed at the DNA Sequencing and Analysis Facility of the University of Minnesota. One sequencing run (one full plate) was performed on a 454 GS FLX system (454 Life Sciences, Branford, CT).

5.3.2 *De novo* assembly

Assembly of reads into contigs was done using the GS De Novo Assembler v1.1.03.24 (454 Life Sciences, Branford, CT) with default parameters (seed step 12; seed length 16; minimum overlap length 40; minimum overlap identity 90%). Contigs smaller than 100 base pairs (bp) were not considered. Large contigs were defined as those with a minimum length of 500 bp.

5.3.3 Alignment

Two types of alignment were performed:

1. Reads were mapped to reference genomes, using the 454 GS Reference Mapper v1.1.03.24 (454 Life Sciences, Branford, CT) with default parameters (seed step 12; seed length 16; minimum overlap length 40; minimum overlap identity 90%). The output of the 454 GS Reference Mapper consists of contigs generated from the mapping of the reads to the reference genomes (mapping

contigs). Similarly to the case of *de novo* assembly, contigs smaller than 100 bp were not considered and large contigs were defined as those over 500 bp.

2. The contigs resulting from *de novo* assembly (assembled contigs) were mapped to reference genomes, using MUMmer version 3.20 (Delcher, Kasif et al. 1999; Delcher, Phillippy et al. 2002; Kurtz, Phillippy et al. 2004). The programs NUCmer and PROmer, which are part of MUMmer were used for mapping. NUCmer performs alignment of DNA, whereas PROmer performs alignments in the six frame amino acid translation of the input DNA sequences. Thus PROmer can detect more alignments than NUCmer since protein sequences are more highly conserved as compared to DNA sequence.

5.3.4 Reference genomes

The reference genomes used in this work are listed in Table 5.1. All reference genomes were originally downloaded from Genbank (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/) on April 2008. Updated files were downloaded from Genbank on June 2009.

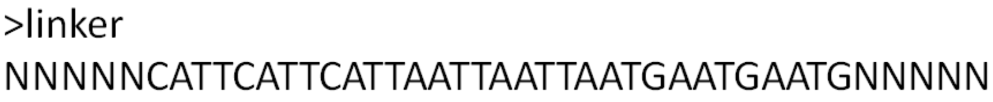
Table 5.1. Genomic features of reference genomes.

Reference genome	**Chromosome sequences**	**Plasmid sequences**	**Total length (Mb)**
Agrobacterium radiobacter K84	2	3	7.3
Agrobacterium tumefaciens C58	2	2	5.7
Agrobacterium vitis S4	2	5	6.3
Rhizobium etli CFN_42	1	6	6.5
Rhizobium etli CIAT 652	1	3	6.4
Rhizobium leguminosarum bv. trifolii WSM2304	1	4	6.9
Rhizobium leguminosarum bv. viciae 3841	1	6	7.8
Rhizobium NGR234	1	2	6.9
Sinorhizobium meliloti	1	2	6.7
Sinorhizobium medicae WSM419	1	3	6.8

5.3.5 Open Reading Frame (ORF) prediction

A pseudochromosome was created by joining the large contigs (in random order) with a linker sequence containing start and stop codons in all six reading frames (Tettelin, Masignani et al. 2005; Wackett, Frias et al. 2007). The linker sequence is provided in Figure 5.4 in FASTA format.

ORF prediction on the pseudochromosome was done using Glimmer and GeneMark. For Glimmer, NCBI Glimmer (ver. 3.02; iterated) with genetic code 11 (Bacteria, Archaea) was used (Delcher, Harmon et al. 1999; Delcher, Bratke et al. 2007). For GeneMark, the web version available at (http://exon.biology.gatech.edu/genemarks.cgi) was used (Borodovsky and McIninch 1993; Besemer and Borodovsky 2005). The final list of ORFs was obtained by combining the prediction of Glimmer and GeneMark. When both programs predicted the same ORF but with different start sites, the longer ORF was kept.



```
>linker
NNNNNCATTCATTCATTAATTAATTAATGAATGAATGNNNNN
```

Figure 5.4. Linker sequence in FASTA format.

5.3.6 Annotation

Genes were annotated by BLAST comparisons to *Agrobacterium tumefaciens* C58 and the databases uniref90 and nt. The comparisons were done using TimeLogic Decypher version 7.6.0 (Active Motif, Inc., Carlsbad, CA), with a significance of 10^{-10} used as threshold. The best reciprocal hit was selected manually. The uniref90 database was downloaded on 09/02/08 from Uniprot (ftp://ftp.uniprot.org/pub/databases/uniprot/uniref/uniref90/). The nt database was downloaded on 02/20/2008 from NCBI (ftp://ftp.ncbi.nih.gov/blast/db/). A significance of 10^{-4} was used as threshold in the case of comparisons against the uniref 90 and nt databases. The output of the search against uniref90 and nt was parsed with a script to select the best hit from a list of close relatives. This script (top20.pl) was provided by Dr. Anne Ruffing.

The predicted ORFs were also submitted to IMG/ER version 3.0 (Markowitz, Ivanova et al. 2008; Markowitz, Szeto et al. 2008) for automatic annotation. Genome visualization was done within IMG/ER and on Artemis release 12 (Rutherford, Parkhill et al. 2000; Berriman and Rutherford 2003) run locally. Genbank files were created for use within Artemis.

5.4 Results

5.4.1 Sequencing output

The output of the sequencing run was a total of 399,219 reads and 92,994,272 bases. Read length distribution was wide, ranging from 15 bp to 367 bp. The median read length was 246 bp. The read length distribution can be seen in Figure 5.5.

The quality average, a Phred equivalent, was 33. A Phred quality score of 30 indicates that the probability of an incorrect base call is 1 in 1,000, or a base call accuracy of 99.9%. A Phred quality score of 40 indicates that the probability of an incorrect base call is 1 in 10,000 or a base call accuracy of 99.99%.

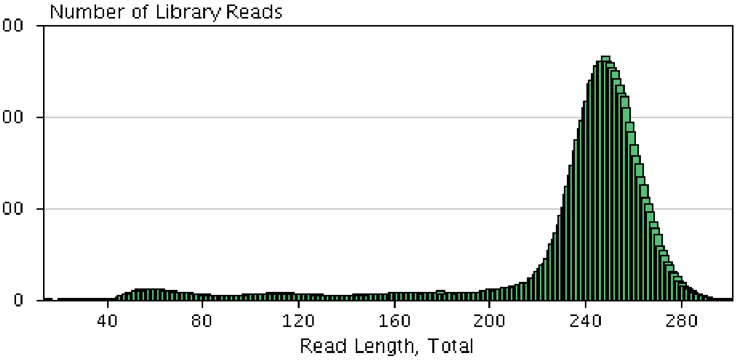


Figure 5.5. Read length distribution of sequencing run using GS FLX system.

5.4.2 *De novo* assembly generates 95 large contigs

Of the 399,219 reads 397,209 were assembled into contigs using the GS De Novo Assembler. This represents 99.50% of the total reads. The vast majority of reads (98.24%) were completely assembled into contigs. A small fraction of reads (1.09%) were partially assembled. A positive linear correlation between reads assembled into a contig and its length was observed (Figure 5.6), that is, the more reads assembled into a contig resulted in longer contigs.

De novo assembly resulted in 176 contigs, 95 of them being longer than 500 bp. The largest contig was 364,263 bp. The average contig size and N50 for both all contigs and large contigs (those ≥ 500 bp) can be found in Table 5.2. N50 is the length of the smallest contig in the set with additive lengths that represent at least 50% of the assembly, when the contig set is ordered from largest to smallest (Miller, Koren et al.

2010). For the resulting genome length of 5.5 Mb, the total sequenced bases (93 Mb) corresponded to a 17x coverage.

Table 5.2. Features of assembled contigs.

Feature	All contigs (≥ 100 bp)	Large contigs (≥ 500 bp)
Number of contigs	176	95
Average contig size (bp)	31,122	57,493
N50 (bp)	122,028	122,028
Total length (bp)	5,477,603	5,461,917

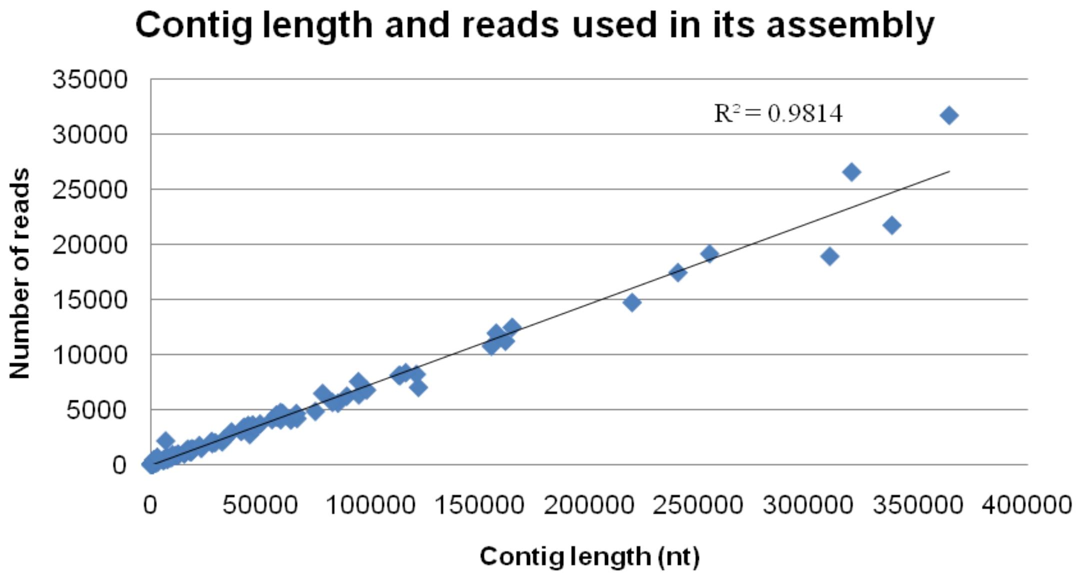


Figure 5.6. Correlation between contig length and number of reads assembled into contigs.

5.4.3 Read alignment to reference genomes reveals that *Agrobacterium* sp. ATCC 31749 is most similar to *Agrobacterium tumefaciens* C58

When the genome of a close relative is available, reads can be aligned to reference genomes to guide the assembly. From preliminary results (not shown) using 16S rRNA, it was known that *Agrobacterium* sp. ATCC 31749 is a close relative to the phytopathogenic *Agrobacterium tumefaciens* C58. Since the taxonomy of the *Agrobacterium* genera has been under review and its distinction from the *Rhizobium* genera questioned, read alignment was performed to each of the so far sequenced members of *Agrobacterium*, *Rhizobium*, and *Sinorhizobium* (Listed in Table 5.1).

General results of aligning the reads to other Rhizobia genomes using the GS Reference Mapper (454 Life Sciences, Branford, CT) are given in Table 5.3. The results indicate that *Agrobacterium* sp. ATCC 31749 shows the highest similarity to *Agrobacterium tumefaciens* C58, and that similarity extends over large areas of the genome. More than 345,000 reads (87%) were mapped to the genome of *Agrobacterium tumefaciens* C58, whereas only 50,000-80,000 (13-20%) reads were mapped to the other genomes.

The GS Reference Mapper uses the aligned reads to guide assembly. The assembly guided by the reference genome of *Agrobacterium tumefaciens* C58 resulted in the less number of contigs, 353. The total length of the contigs resulting from this guided assembly was 4.7 Mb. In contrast, the assemblies guided by the rest of the reference genomes resulted in a large number (1000-1800) of short contigs. Whereas the longest contig obtained by using *Agrobacterium tumefaciens* C58 as a guide was 141 Kb, for the rest of the reference genomes it was only in the 1-2 Kb range. This reveals that the genome of *Agrobacterium* sp. ATCC 31749 presents long syntenic regions with the genome of *Agrobacterium tumefaciens* C58, but not with the genome of other Rhizobia.

Although 87% of the sequencing reads aligned to the reference genome of *Agrobacterium tumefaciens* C58, the reads did not aligned to all parts of the four genomic elements of *Agrobacterium tumefaciens* C58. As can be seen in Table 5.4, while most of the circular and linear chromosomes were mapped by our reads, only one third of the plasmid At was mapped, and a minimal portion of the plasmid Ti was mapped. The genes responsible for virulence in *Agrobacterium tumefaciens* C58 are localized to plasmid Ti. This confirms the phenotype of *Agrobacterium* sp. ATCC 31749, which is not phytopathogenic.

Table 5.3. Assembly of *Agrobacterium* sp. ATCC 31749 reads guided by the genome sequence of other Rhizobia.

Reference genome	Reads mapped	% reads mapped	Contigs	Large contigs	Largest contig size (bp)	Total contig length (Kb)
Agrobacterium tumefaciens C58	347,497	87	353	286	141,213	4,756
Agrobacterium radiobacter K84	71,539	18	1,599	51	2,022	304
Agrobacterium vitis S4	58,008	15	1,172	26	1,165	215
Rhizobium etli CFN_42	72,847	18	1,572	41	2,419	303
Rhizobium etli CIAT 652	75,732	19	1,714	55	2,521	331
Rhizobium leguminosarum bv. trifolii WSM2304	78,765	20	1,820	46	1,624	346
Rhizobium leguminosarum bv. viciae 3841	78,942	20	1,788	49	1,357	343
Rhizobium NGR234	61,510	15	1,336	23	1,197	246
Sinorhizobium meliloti	57,788	14	1,196	23	1,040	213
Sinorhizobium medicae WSM419	50,291	13	930	15	1,037	161

Table 5.4. Mapping of reads to the four genetic elements of *Agrobacterium tumefaciens* C58.

Element	Size (Kb)	% Covered
Circular chromosome	2,882	96.95
Linear chromosome	2,105	87.91
Plasmid At	551	32.74
Plasmid Ti	217	1.86

5.4.4 Contig alignment to reference genomes confirms similarity between *Agrobacterium* sp. ATCC 31749 and *Agrobacterium tumefaciens* C58

Contigs resulting from *de novo* assembly were aligned to the genomes of other Rhizobia for their comparison. The comparative analysis can provide information on conserved and deleted regions, and rearrangements. The alignment of the 176 contigs obtained from the assembly (assembled contigs) was done using PROmer. The output was filtered for regions with a minimum percent identity of 50%. PROmer performs alignment after performing six frame amino acid translation of the DNA sequences. Table 5.5 shows a summary of the contig alignment results.

Table 5.5. Mapping of assembled contigs to reference genomes using PROmer.

Reference genome	Total alignment length (Mb)	% of reference genome mapped
Agrobacterium tumefaciens C58	4.9	87
Agrobacterium radiobacter K84	3.1	43
Agrobacterium vitis S4	2.9	46
Rhizobium etli CFN_42	3.0	46
Rhizobium etli CIAT 652	3.0	46
Rhizobium leguminosarum bv. trifolii WSM2304	3.1	45
Rhizobium leguminosarum bv. viciae 3841	3.3	42
Rhizobium NGR234	2.9	42
Sinorhizobium meliloti	2.9	43
Sinorhizobium medicae WSM419	2.9	42

From the results in Table 5.5 it is interesting to note that the contigs generated by *de novo* assembly map most reference genomes by a very similar length (~ 3.0 Mb). The exception was *Agrobacterium tumefaciens* C58, which had a larger length mapped (4.9 Mb corresponding to 87%). Of the 176 assembled contigs, 67 aligned to all reference genomes. The similarity in the total length mapped by the same contigs seems to define the core genome for the Rizobia. All chromosome and plasmid elements showed some region mapped by the assembled contigs (even if small). The exception was plasmid pAgK84 of *Agrobacterium radiobacter* K84, which did not show any alignment to the assembled contigs.

Of the 81 contigs that did not show any alignment, almost all of them are short contigs, less than 250 bp in length. There were only three contigs longer than 250 bp that did not show alignments (contig00044, contig00071, and contig00085). Of the three, only one of them (contig00044) is of a significant length (18 Kb) to represent a completely new region in *Agrobacterium* sp. ATCC 31749. The other two contigs, contig00085 and contig00071, are 821 and 1,158 bp, respectively. At this length they could still code for a complete or partial ORF. Alignment to the nt database using BLAST revealed that contig00085 hits 820 bp of the transposase gene ISRpe1 of *Rickettsia peacockii* with 100% identity. Its appearance in a single contig could indicate posible contamination. Alignment to the nt database using BLAST did not result in significant hits for contig00071.

Alignment of the assembled contigs to the *Agrobacterium tumefaciens* C58 genome was done at the nucleotide level using NUCmer. Using NUCmer resulted in 78 assembled contigs aligned to the reference genome. All of these contigs were also aligned to the reference genome when PROmer was used. Eight contigs were aligned to the reference genome by PROmer only. These eight contigs (Table 5.6) could represent regions that are more divergent between the two genomes.

Table 5.6. Contigs mapped to *Agrobacterium tumefaciens* C58 only when PROmer was used.

Contig	Length (bp)	Alignment element	Hits	% Cov Ref	% Cov Qry
contig00021	1053	Linear chromosome	Single	0.05	99.43
contig00023	2858	Linear chromosome	Single	0.08	55.46
contig00024	46339	Plasmid Ti	Multiple	3.12	14.44
contig00024	46339	Plasmid At	Multiple	1.24	14.47
contig00145	44277	Linear chromosome	Single	0.03	1.46
contig00145	44277	Plasmid At	Single	0.47	5.82
contig00161	2037	Linear chromosome	Single	0.05	46.69
contig00162	677	Plastmid At	Single	0.04	37.22
contig00165	2383	Plasmid Ti	Multiple	0.87	75.66
contig00169	5455	Linear chromosome	Single	0.04	14.41

5.4.5 ORF prediction and annotation

By combining the ORF predictions obtained from Glimmer and GeneMark, a final list of 5585 ORFs was obtained. When the same ORF was predicted in both programs but with a different start site, the longest ORF was retained. Most of the ORFs were predicted by both programs. Up to 75% of the ORFs were predicted by both programs with exactly the same length, 16% were predicted with both programs but with different start codons, and only 9% of the ORFs were predicted by only one of the two programs (Table 5.7).

Table 5.7. Source of predicted ORFs.

Method	Number of ORFs
Both programs, same length	4202
GeneMark longer ORF than Glimmer	539
Glimmer longer ORF than GeneMark	330
GeneMark only	180
Glimmer only	334
TOTAL	**5585**

Annotation of the predicted ORFs was done using multiple sources. BLAST searches were done against the genes of *Agrobacterium tumefaciens* C58, and the databases uniprot 90 and nt. The predicted ORFs were also submitted to IMG/ER for automatic annotation. Table 5.8. summarizes the results of annotation using the different resources. In the case of annotation based on hits to *Agrobacterium tumefaciens* C58 ORFs, BLAST results were checked manually to select the best reciprocal hit. In Table 5.8, ambiguous hit refers to cases in which the predicted ORF from *Agrobacterium* sp. ATCC 31749 had an equally good hit to more than one ORF from *Agrobacterium tumefaciens* C58 or was completely within another ORF with longer sequence.

Table 5.8. Summary of ORFs with hits in the different databases.

Reference system	Hit	No hit	Ambiguous hit
Agrobacterium tumefaciens C58	4277 (77%)	1002 (18%)	306 (5%)
Uniprot90	5464 (98%)	121 (2%)	
nt	5064 (91%)	521(9%)	
IMG	4286 (77%)	1299 (23%)	

5.5 Discussion

5.5.1 Agrobacterium sp. ATCC 31749 is most similar to *Agrobacterium tumefaciens*

Mapping of sequencing reads to reference genomes of several Rhizobia revealed that only in the case of *Agrobacterium tumefaciens* C58 there was a significant number of reads mapped (87%, see Table 5.3). This confirmed preliminary results obtained by Dr. Anne Ruffing at the Georgia Institute of Technology from 16S rRNA analysis, which indicated that *Agrobacterium* sp. ATCC 31749 was most similar to *Agrobacterium tumefaciens* C58.

The GS Reference Mapper was able to guide the assembly of reads, by using the *Agrobacterium tumefaciens* C58 genome as a reference, into 353 contigs with a total length of 4.8 Mb. The contig size reveals that both strains have extensive regions of similarity, 81% of the 353 resulting contigs are longer than 500 bp, and the largest contig has a length of 141Kb. *Agrobacterium tumefaciens* C58, although known to have orthologous genes to those required for the production of curdlan, has not been reported to produce the exopolysaccharide. Another clear difference between the two strains is that *Agrobacterium tumefaciens* is phytopathogenic, whereas *Agrobacterium* sp. ATCC 31749 is not.

The circular chromosome of *Agrobacterium tumefaciens* C58, which contains more housekeeping genes than the linear chromosome (Goodner, Hinkle et al. 2001; Wood, Setubal et al. 2001), was almost completely mapped (97%) by the reads obtained from *Agrobacterium* sp. ATCC 31749 sequencing (Table 5.4). Approximately 88% of the linear chromosome of*Agrobacterium tumefaciens* C58 was mapped by our reads. In the case of the At plasmid the mapping was reduced to only 33% and was minimal (2%) for plasmid Ti. None of the virulence genes (*vir*), located in plasmid Ti, were found in the sequence from *Agrobacterium* sp. ATCC 31749, which confirms its non phytopathogenicity. Of the genes involved in nodulation, two (*nodL* and *nodT*) are located in the circular chromosome of *Agrobacterium tumefaciens* C58, and three (*nodN*, *nodX*, and *nodW*) in the linear chromosome. *nodL*, *nodN*, and *nodW* are present in the genome of *Agrobacterium* sp. ATCC 31749, *nodX* is not, and *nodT* gave ambiguous hits.

5.5.2 Genes relevant to curdlan production

The curdlan synthesis gene (*crdS*) ortholog is located in the linear chromosome in *Agrobacterium tumefaciens* C58. The nucleotide sequence for *crdS* on *Agrobacterium* sp. ATCC 31749 is the only sequence publicly available from the *crdASC* operon. The sequence of *crdS* for *Agrobacterium* sp. ATCC 31749 obtained in this study was compared to the publicly available *crdS* sequence and also to the *crdS* ortholog of *Agrobacterium tumefaciens* C58.

A pss_{AG} ortholog is located in the circular chromosome in *Agrobacterium tumefaciens* C58. pss_{AG} is required for maximal curdlan yields. The protein sequence corresponding to pss_{AG} from *Agrobacterium* sp. ATCC 31749 is publicly available. We compared the translated ORFs from *Agrobacterium tumefaciens* C58 and *Agrobacterium* sp. ATCC 31749 to the deposited protein sequence.

Other genes relevant to curdlan production include the two-component system *ntrBC*. Orthologues of *ntrBC* are located in the circular chromosome of *Agrobacterium tumefaciens*. *ntr* mutants can not produce curdlan in some media. Since no sequence was previously available for these genes, the orthologues from *Agrobacterium tumefaciens* C58 were compared to the sequence obtained here for *Agrobacterium* sp. ATCC 31749.

Genes *rcdAB*, annotated as curdlan synthase and curdlan synthesis protein, respectively, are located in the *Agrobacterium tumefaciens* C58 plasmid At. These genes are also present in the second chromosome of *Agrobacterium radiobacter*, a species known to produce curdlan. These genes, however, were not present in the *Agrobacterium* sp. ATCC 31749 genome, which mapped only 33% of the sequence of plasmid At. The absence of *rcdAB* in *Agrobacterium* sp. ATCC 31749 suggests that these genes are not required for curdlan production.

Table 5.9 summarizes the sequences available for each source, and Table 5.10 shows the results of nucleotide sequence alignment. From Table 5.10 it is clear that even though the percent identity is very high for all sequences compared, it is not 100%. Since the differences in sequence could translate into different amino acids, the translated sequences were compared to determine the number and type of substitutions (conserved, semi-conserved, or not conserved) (Table 5.11).

The comparison of CrdS revealed that not only is there a difference between *Agrobacterium tumefaciens* C58 and the sequenced *Agrobacterium* sp. ATCC 31749, but that the sequence from this work has differences with respect to the previously published CrdS sequence. The genes flanking *crdS* also present differences at the amino acid level. These differences could explain why *Agrobacterium tumefaciens* does not produce curdlan. However, at the coverage level obtained in this project (17x) it is also possible that the differences are the result of sequencing errors or missassemblies. Further experimentation is required to confirm these differences.

Results for the phosphotidyl serine synthase revealed that the translated protein from *Agrobacterium tumefaciens* C58 and *Agrobacterium* sp. ATCC 31749 sequenced in this work, are longer than the available *pss*$_{AG}$ sequence. It is possible that the previously reported *pss*$_{AG}$ sequence is only a partial sequence. Finally, the two-component system NtrBC does not present differences at the amino acid level between *Agrobacterium tumefaciens* C58 and *Agrobacterium* sp. ATCC 31749.

Although a clear difference that could explain the lack of curdlan production in *Agrobacterium tumefaciens* C58 was not yet clearly determined, these results suggest candidates for further investigation.

Table 5.9. Genes relevant to curdlan production.

Data Type	Length	Agrobacterium tumefaciens C58 locus	*Agrobacterium* sp. ATCC 31749 locus (this study)	*Agrobacterium* sp. ATCC 31749 locus (publicly available)
Nucleotide	1266 bp	Atu3055	r1842	
Nucleotide	1965 bp	Atu3056	r1841	*crdS* (AF057142)
Nucleotide	1458 bp	Atu3057	r1840	
Protein	274 aa	Atu1062	r3901	*pssAG* (AAL01116)
Nucleotide	1149 bp	Atu1445 (*ntrB*)	r4548	
Nucleotide	1452 bp	Atu1446 (*ntrC*)	r4547	

Table 5.10. Comparison of *crdASC* elements at the nucleotide level.

Nt comparison	Alignment length	Mismatches	Gaps	% identity
Atu3055 – r1842	1266	23	0	98.18
Atu3056 - r1841	1965	23	0	98.83
Atu3057 – r1840	1458	14	0	99.04
Atu3056 - *crdS*	1965	33	0	98.32
r1841 - *crdS*	1965	10	0	99.49

Table 5.11. Amino acid level comparison of proteins relevant to curdlan production.

Comparison	Substitutions	Semi-conserved substitutions	Conserved substitutions	Offset
Atu3056 – CrdS	4	5	4	0
r1841 - CrdS	3	2	3	0
Atu3056 – r1841	1	3	1	0
Atu3055 – r1842	3	1	3	0
Atu3057 – r1840	1	1	0	0
Atu1062 - Pss$_{AG}$	1	1	0	14
r3901 - Pss$_{AG}$	1	1	0	16
Atu1062 – r3901	1	1	2	2
Atu1445 – r4548	0	0	0	0
Atu1446 – r4547	0	0	0	0

Other genes of particular interest were the 1002 ORFs from *Agrobacterium* sp. ATCC 31749 with no homologue in *Agrobacterium tumefaciens* C58. 116 of these ORFs could not be annotated. Of the remaining 886 ORFs, 257 were annotated as putative uncharacterized proteins, reducing the number of genes with meaningful annotations to 629. Of these 629 genes, seventeen corresponded to ABC transporters, and six to probable sugar ABC transporters. Twenty-six family-specific transcriptional regulators were also detected.

5.5.3 Transcriptome analysis of curdlan production

The draft genome of *Agrobacterium* sp. ATCC 31749 was used by our collaborators at the Georgia Institute of Technology to design a custom oligonucleotide microarray containing probes for 5580 ORFs. The microarray was used to compare two stages of growth: the nitrogen-rich growth stage and the nitrogen-limited curdlan production stage. Several regulators and stress response genes were found up-regulated in the nitrogen-limited stage. Among the down-regulated genes in the nitrogen-limited stage are transporters and other genes involved in metabolism. Genes in the *crdSAC* operon were found upregulated by 60-fold in the nitrogen-limited stage. Some of the differentially expressed genes corresponding to regulators were selected for creating knockout mutants for further study.

5.6 Concluding remarks

Agrobacterium sp. ATCC 31749 is a producer of the exopolysaccharide curdlan, which has uses in the food industry as a gelling agent. Due to the lack of genetic information on this organism, metabolic engineering attempts were limited. Genome sequencing of this organism had two main purposes: gain information on genes potentially involved in curdlan production, and compare the genome of *Agrobacterium* sp. ATCC 31749 to other Rhizobia genomes, as taxonomic classifications are under review.

The reads obtained from a run of sequencing using were *de novo* assembled into contigs. The reads were also mapped to reference genomes to guide assembly. *De novo* ssembled contigs were also aligned to the genomes of other Rhizobia. The contigs were randomly joined into a pseudochromosome which was used for ORF prediction.

The ORFs were annotated based on hits to *Agrobacterium tumefaciens* C58, the uniref90 and nt databases, and using the IMG/ER system.

The resulting *Agrobacterium* sp. ATCC 31749 draft genome was compared to other Rhizobia genomes and particular focus was given to genes known to be involved in curdlan production regulation. The draft genome was used by our collaborators at the Georgia Institute of Technology to create a microarray with which further studies are being conducted in an attempt to identify genes affecting the regulation of curdlan production. Key regulators potentially involved in curdlan production regulation were selected for further study using knock-out mutants.

Chapter 6
Genome-wide mutation profiling of a clavulanic acid high producer

6.1 Summary

Streptomyces are well known as secondary metabolites producers. *Streptomyces clavuligerus* is one of the industrially important organisms in this genus. It produces two β-lactam antibiotics: cephamycin C and clavulanic acid. The clusters for the production of these two antibiotics are located next to each other in the chromosome.

Industrial screening programs have resulted in high producer strains of clavulanic acid, which is used commercially in combination with amoxicillin in the GlaxoSmithKline product Augmentin^{TM}. In this project we obtained a clavulanic acid high producer strain from GlaxoSmithKline and performed genome-wide mutation screening using next generation sequencing data. The strain was compared to the WT strain which was also sequenced in this work and to a reference genome recently available.

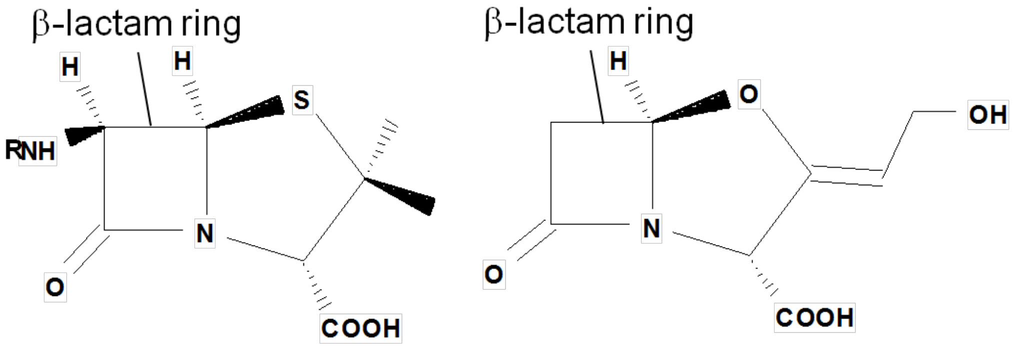
6.2 Introduction

Actinomycetes produce two-thirds of the known antibiotics, and of those 80% are produced by *Streptomyces* (Kieser, Bibb et al. 2000). *Streptomyces* are Gram-positive bacteria with high G+C content. They undergo a complex differentiation cycle, which includes the development of vegetative and aerial mycelia, and the formation of spores. Production of secondary metabolites is highly regulated and involves pathway specific regulators and pleiotropic regulators.

β-lactam antibiotics, of which penicillin is the best known, are characterized by the presence of the four-membered β-lactam ring (Figure 6.1). In addition to penicillin, other β-lactams include cephalosporins, monobactams, and carbapenems. β-lactam antibiotics are produced by filamentous fungi, Gram-negative, and Gram-positive bacteria (Brakhage, Al-Abdallah et al. 2005).

Streptomyces clavuligerus produces antibiotics belonging to two major groups of β-lactam compounds: the sulfur-containing cephalosporins and the oxygen-containing clavams (Paradkar and Jensen 1995). Most of the clavams produced by *Streptomyces clavuligerus* display a (3*S*, 5*S*) stereochemistry, except clavulanic acid which displays a (3*R*, 5*R*) conformation (Jensen, Paradkar et al. 2004). *Streptomyces clavuligerus* has been used commercially to produce clavulanic acid (Figure 6.1). Clavulanic acid was the first clinically useful agent against β-lactamases (Sutherland 1991), enzymes which target the β-lactam ring and render β-lactam antibiotics useless. β-lactamases are greatly responsible for bacterial resistance to β-lactam antibiotics. The β-lactamase inhibitory effect of clavulanic acid seems to be linked to its particular stereochemistry, as none of the other clavams exhibit this inhibitory effect. Nonetheless, some of the clavams have an antibacterial effect.

Cephamycin C and clavulanic acid production occur in parallel. The biosynthetic genes for both β-lactams are located in adjacent clusters, forming a β-lactam supercluster in the genome of *Streptomyces clavuligerus*.



Penicillin

Clavulanic Acid

Figure 6.1. Structure of penicillin and clavulanic acid. The compounds have similar structures, including the presence of the □-lactam ring. In the case of penicillin the ring contains sulfur, whereas in clavulanic acid it contains oxygen.

In 1967 large-scale screenings for β-lactamase inhibitors began, as isolates producing β -lactamases had increased considerably and resistance was being passed between bacteria through plasmids, even between different species. Clavulanic acid was identified in 1972. Amoxicillin, co-administered with clavulanic acid was launched as Augmentin^{TM} in 1981. The amoxicillin-clavulanic combination is still of relevance today, in the treatment of community-acquired infections, like respiratory tract infections (Geddes, Klugman et al. 2007).

Strain improvement programs have resulted in clavulanic acid high producing mutants. Conventional strain development programs focused on obtaining the highest producer, but at the time it was not possible to determine the mutations responsible for high productivity. Rational strain improvement through genetic engineering later resulted in further strain improvements and higher clavulanic acid production. In many cases the starting strain for genetic engineering was already a high producing strain, which was identified years earlier by screening. Thus, even when specific mutations were introduced in the strains many other mutations could have already been selected for.

With the availability of next generation sequencing technologies, it is now possible to perform genome-wide comparisons of strains in an affordable manner. In this study, an industrial strain of *Streptomyces clavuligerus* considered a high producer of clavulanic acid is compared to the wild type (WT) strain using next generation sequencing (NGS) technology. Genome-wide mutation analysis was performed to identify the mutations most likely responsible for high production.

6.3 Availability of the *Streptomyces clavuligerus* genome sequence

Two groups were known to be working on sequencing *Streptomyces clavuligerus* when this project started, one at the Korea Research Institute of Bioscience and Biotechnology, KRIBB, and the other at the Broad Institute. At that point, however, no sequence information was available from KRIBB and only a draft sequence was available from Broad. The draft reference from the Broad Institute consisted of 597 contigs with a total length of 6.7 Mb in 158 supercontigs and a total length of 6.9 Mb. This size, however, was small compared to the genomes of other *Streptomyces* already sequenced, which ranged from 8.5 Mb for *Streptomyces avermitilis* to 10.1 Mb for *Streptomyces scabiei*. Thus it was decided to sequence the wild type strain at the same level of coverage as the mutant strain.

Recently, on March 29th 2010, DSM in collaboration with the University of Groningen, in the Netherlands, announced the sequence of a 1.8 Mb megaplasmid of *Streptomyces clavuligerus* which contained an extremely high number of secondary metabolite biosynthetic clusters. In this work by Medema et al. (Medema, Trefzer et al. 2010), the whole genome of *Streptomyces clavuligerus* was sequenced using Sanger shotgun sequencing of three libraries with three different insert sizes. The assembly took advantage of an optical restriction map and finally as many gaps as possible were filled. The genome is not 100% complete, though. It was estimated to represent 99.7% of it. This genome, further referred as DSM reference, consists of a chromosome of 6.8 Mb and a linear megaplasmid of 1.8 Mb. In total the assembly consists of 279 contigs, 241 of which correspond to the chromosome and 38 to the plasmid.

6.4 Materials and methods

6.4.1 Strains, media, and growth

Streptomyces clavuligerus NRRL 3585 was used as the wild type strain. *Streptomyces clavuligerus* strain X from GlaxoSmithKline, obtained through collaboration with Prof. David Sherman at the University of Michigan, was used as the mutant, high producer strain.

Trypticase soy broth supplemented with 1% soluble starch (Sup. TSB) was used for seed cultures. For genomic DNA (gDNA) extraction, *Streptomyces clavuligerus* was cultivated in chemically defined media (Table 6.1). For clavulanic acid production studies, fermentations were carried out in soy-base media, SP, (Table 6.3). pH was adjusted to 6.80 for all media.

Table 6.1. Aharonowitz and Demain modified media (Aharonowitz and Demain 1979).

Component	Amount
Glycerol	10 g
L-asparagine	2 g
K_2HPO_4	3.5 g
$MgSO_4*7H_2O$	1.23 g
MOPS	20.9 g
Bacto yeast extract	1 g
NH_4Cl	1 g
Trace salt solution (Table 6.2)	1 ml
Water	To 1 L

Table 6.2. Trace salt solution for Aharonowitz and Demain modified media.

Component	Amount
$FeSO_4*7H_2O$	0.1 g
$MnCl_2*4H_2O$	0.1 g
$ZnSO_4*7H_2O$	0.1 g
$CaCl_2$	0.1 g
Water	100 ml

Table 6.3. Soy-based media, SP (Paradkar and Jensen 1995).

Component	Amount
Soybean flour	15 g
Soluble starch	4.7 g
KH_2PO_4	0.1 g
$FeSO_4*7H_2O$	0.2 g
Water	To 1 L

For gDNA extraction, cultures with a 5 ml working volume were inoculated with spores (and mycelia) scrapped from a plate. Cultures for pre-inoculum were incubated for approximately 48 hours and then used to inoculate a 50 ml seed culture in a 250 ml baffled flask. The seed culture was maintained for approximately 36 hours, washed twice with sterile water and then used to inoculate the main culture. The main culture consisted of 400 ml working volume in 2L baffled flasks. Main cultures were maintained for approximately 36 hours. All cultures were incubated at 28°C in a MaxQ 5000 orbital shaker (Barnstead International, Thermo Scientific, Asheville, NC) at 220 rpm. Culture samples were kept at -20°C until gDNA extraction was performed.

For clavulanic acid studies a similar strategy was used, except that the seed cultures were either used directly to inoculate the main fermentor or used to prepare aliquots consisting of one volume of culture mixed with one volume of 40% sterile glycerol. These aliquots were kept at -80°C until use. The main fermentor was inoculated to an initial optical density at 595 nm in the range of 4-7. As clavulanic acid is unstable at basic pH, the pH of the cultures was manually monitored and controlled to a pH of 6.8 with 1N HCl.

Growth was monitored by optical density measurement at 595 nm (OD_{595}) with a UV-160 spectrophotometer (Shimadzu, Columbia, MD). Dilutions were done as necessary to obtain measurements below 1.2, which is the linear measurement limit.

6.4.2 Clavulanic acid determination

Clavulanic acid was measured by high-pressure liquid chromatography (HPLC). One volume of culture supernatant was mixed with five volumes of imidazole reagent and kept at room temperature for 15 minutes. The imidazole reagent consisted of 8.25 g of imidazole dissolved in 100 ml of water. The pH was adjusted to 6.8 with HCl (Bird, Bellis et al. 1982). HPLC determination of clavulanic acid was done as described by

(Foulstone and Reading 1982), except that the pH of the buffer was adjusted to 3.7 (Paradkar, Aidoo et al. 1998) and detection was done at 312 nm (Liras and Martin 2005). Basically, derivatized samples (50 μl) were analyzed on a System Gold equipment consisting of 126 Solvent Module, 508 Autosampler, and 166 Programmable Detector module (Beckman Coulter, Brea, CA) with a Ultrasphere ODS C18 Reversed Phase Column 235330 (Beckman Coulter, Brea, CA), with an isocratic method with a flow of 1 ml/min using a buffer consisting of 0.1 M KH_2PO_4-6% methanol. Under these conditions pure clavulanic acid had a retention time of 6.7 min. Pure clavulanic acid was used to construct a calibration curve. Pure clavulanic acid was a kind gift from Prof. Susan Jensen at the University of Alberta.

6.4.3 gDNA and RNA extraction

gDNA extraction was done using the Kirby-mix method as described by (Kieser, Bibb et al. 2000). Briefly, lysis with lysozyme is followed by lysis with the phenolic detergent Kirby-mix, and phenol-chloroform extraction. gDNA is precipitated with isopropanol, spooled and redissolved in TE buffer. This procedure was followed with further purification using the QIAGEN Genomic-tip 500/G (Qiagen, Valencia, CA).

RNA extraction was done using the RNeasy Mini Kit (Qiagen, Valencia, CA). Cell pellets kept at -80°C were deposited in mortars, immersed in liquid nitrogen and grinded completely. Further isolation continued with the RNeasy Mini Kit as described in the manufacturer's manual. The mortars were cleaned by immersion in 3% H_2O_2, rinsed with double autoclaved DI-UV water, baked overnight at 180°C, and kept at -80°C. Removal of residual gDNA was done with the Turbo-DNA free kit (Ambion, Applied Biosystems by Life Technologies, Carlsbad, CA), following the manufacture's protocol for rigorous DNase treatment.

6.4.4 cDNA synthesis

RNA was converted to cDNA for use in the measurement of transcript levels. Up to 2 μg of total RNA were mixed with 1 μg of random hexamers in a 15 μl reaction. The mixture was incubated at 70°C for 10 minutes and chilled on ice for 5 minutes. 14.5 μl of RT cocktail (Table 6.4), and 0.5 μl of Superscript III were added. To prepare the No-RT control 0.5 μl of water was added instead of Superscript III. The mixture was incubated at 25°C for 10 minutes, followed by incubation at 50°C for 60 minutes, and at 72°C for 5 minutes. The cDNA synthesized was directly used in qRT-PCR (following section), or stored at -20°C until use. Superscript III, 5x RT buffer, and DTT, and dNTPs were from Invitrogen (Life Technologies, Carsbad, CA).

Table 6.4. RT cocktail composition

Component	Amount
5x RT buffer	6 μl
0.1 M DTT	3 μl
dNTP mix (10 mM dGTP, 10 mM dCTP, 4 mM dTTP, 4 mM dATP)	1 μl
RNase-free water	4.5 μl

6.4.5 qRT-PCR

Quantitative Real Time Polymerase Chain Reaction was used to measure the gDNA and transcript levels using a Mx 3000P instrument (Stratagene, Agilent Technologies, Santa Clara, CA) with MxPro APCR Software version 3.20. Each reaction consisted of 6.25 μl of 2x SYBR green mix, 0.19 μl of ROX reference dye (diluted 500 fold), 4.05 μl of RNase free water, 0.5 μl of each primer (20 nM), and 1 μl of template at a concentration of 5 ng/μl (unless otherwise noted). The 2x SYBR green mix and ROX reference dye are part of the Brilliant SYBR Green QPCR Master Mix (Stratagene, Agilent Technologies, Santa Clara, CA). Each gene was tested in triplicate. When transcript levels were tested a no reverse transcription (No RT) control was also included. The following program was used: 95°C for 10 minutes, followed by 40 cycles at 95°C for 30 seconds, 55°C for one minute, and 72°C for 30 seconds followed by a dissociation protocol to obtain a melting curve. Table 6.5 lists the primers used in this work.

Table 6.5. Sequence of primers used in qRT-PCR

Locus	Left primer	Right primer	Experiment
SCLAV_0244	AGTGGTTCCAGGAGATCGAG	CAGCGGGAACTCACTGACC	Genome coverage
SCLAV_1267	AAGAAGGCGAAGAGGCAGAG	GGCGACCAGGGTCTTGAA	Genome coverage
SCLAV_1555	GTGATCTGTCTGCTGCCGTA	CCCAGACCGGAGTTCTTGTA	Transcript measurement
SCLAV_1556	TCCATCTGGAGGAGAACCAC	GTTCCAGCGAACGGAGATAG	Transcript measurement
SCLAV_1957	CTCCTGGAGACGTCGGACTA	GTACCACCGAGTCCACCAGT	Transcript measurement
SCLAV_2321	CCTTCCTGAAGTTCCTGCTG	GTCCGTCTGGTCGGTGTAGT	Transcript measurement
SCLAV_2335	GTCGAGCTGTCCCAGATCAT	AGATTGGACCGGATGAACAC	Transcript measurement
SCLAV_2335	GTGTTCATCCGGTCCAATCT	GGCGTAGGTGGAGAACTTGA	Genome coverage
SCLAV_2917	GGCTGCTCTTCTTCACCAAC	GGCGAGTCGTAGTCCTTCAG	Control
SCLAV_2925	AGGTCAAACCGCTCTACGG	GAAGTTGTCGACGATCAGCA	Transcript measurement
SCLAV_3625	AAGAAGGTCACGGGGCTTAT	GTGAAGGAACGGTCCTCGTA	Transcript measurement
SCLAV_3643	AAGGGCATCAAGATCCAGTG	GTTCTTGACATCGCCCTTGT	Transcript measurement
SCLAV_4175	CGACGACTACGTCACCAAAC	GCTGACCTCGTGACTCTCCT	Transcript measurement
SCLAV_4181	TCACCTCCAAGGACCTGTTC	GCATCGTCATACAGCTCGAC	Transcript measurement
SCLAV_4190	CTCGACATCCTCGTCAACAA	GAACTTCGTGGCCTGGTAGA	Transcript measurement
SCLAV_4191	GTGGACGATCTCCATCTGCT	GAACTCCGCCTCGTACAGC	Transcript measurement
SCLAV_4194	TGGCCTCTCCGATAGTTGAC	CAGCAGCAGATAACCGTCCT	Transcript measurement
SCLAV_4197	AGTCTGGAGACCGCTCATGT	GACCTCGTCGAAGAGAATCG	Transcript measurement
SCLAV_4204	CAGGTCATCTCCAAGGAAGC	GGAAGACGAAGAGGTCGATG	Transcript measurement
SCLAV_4205	GCGTTCTACCTCGTCGACTC	CGGATACATGGAGTCCTGCT	Transcript measurement
SCLAV_4208	GACCTGTCCCGATGTCAGAT	CGGTACCAGTACGGGATCAT	Transcript measurement
SCLAV_4210	CTTCTTCAGCCATGTCGTCTC	GGGTGGGCTCCAGTTCTT	Transcript measurement
SCLAV_4500	ACCGGTCCTCACCCATCT	CAGGCCATCAGGATCTGC	Transcript measurement
SCLAV_4545	AGCAGTTCAAGGGTGTCGTC	AGTTGTCGCGCTTCTCCTT	Transcript measurement
SCLAV_p0171	CCAGGCACTTCGTCCTATGT	GAGGAAGACCGTCACATCGT	Genome coverage
SCLAV_p0440	CCGGTGAAGGTGTTCTTCTT	GTGTAGTGGTTGGGGTCGTG	Genome coverage
SCLAV_p0448	GGACGTCATCGGTTACCTGT	GTCCTTCAGAGCCTGCAGAT	Genome coverage
SCLAV_p0760	GATGGTCAGAAACGGACACC	GTCGACGAAGTCCACCGTAT	Genome coverage
SCLAV_p0762	CCAGGACCGTGTGATCTACC	CGTTCGACGAAGTACCAGAA	Genome coverage
SCLAV_p0894	ACATCTCACGCCTCCTTCAG	CGTGGTGTTCTTCTTCACCA	Transcript measurement
SCLAV_p1074	AGAGGTCCATCATCCACAGC	GACCAGATCGAGTCGGTGAT	Transcript measurement
SCLAV_p1266	CATCTGTTCATGGGCAACAA	GCCAGAATCGCCTTCTTGTA	Genome coverage
SCLAV_p1269	CCCTGTCCACTTCCGTGAC	GGTCCCTGACACCTCTTTGA	Genome coverage

6.4.6 Sequencing

gDNA library construction and sequencing were done at the National Center for Genome Resources, NCGR (Santa Fe, NM). Sequencing was done using an Illumina (Solexa) Genome Analyzer IIx instrument.

6.4.7 Bioinformatics methods

6.4.7.1 Reference genomes

A reference sequence for *Streptomyces clavuligerus* ATCC 27064 (WT) was downloaded from GenBank, accession number ADGD00000000. Table 6.6 summarizes the characteristics of this reference, which will be further referred to as DSM reference.

Table 6.6. Genome characteristics of *Streptomyces clavuligerus* reference (DSM reference).

	Chromosome	Plasmid
Length (bp)	6,760,392	1,796,500
Coding sequences	5,700	1,581
Secondary metabolite clusters	23	25
Contigs	241	38

6.4.7.2 Read alignment

Alignment of sequenced reads to the DSM reference genome was done using Bowtie version 0.11.3 (Langmead, Trapnell et al. 2009) with the –v mode and allowing for maximum 3 mismatches. Alignment was done treating reads individually or as paired. The output files were obtained in Sequence alignment/Map (SAM) format.

6.4.7.3 Variant determination

SAMtools (Li, Handsaker et al. 2009) version 0.1.7 was used to convert SAM output files to its binary mode, BAM. SAMtools were also used to sort, index, and merge BAM files, and to call single nucleotide differences between the genomes. Genetic variations detected between strains could be supported by only some of the reads covering that position or by 100% of the reads aligning to that region. The term “mutation” will be used to refer to single nucleotide differences supported by 100% of the reads aligning to that position. When the percentage is not 100% the difference will be called variant.

6.4.7.4 Assembly

De novo assembly was done using Velvet version 0.7.49 (Zerbino and Birney 2008). Unpaired and paired assemblies were performed. In both cases a *k*-mer (hash length) of 31 was used. Assemblies were optimized by modifying the coverage cutoff and expected coverage parameters. Initial optimization values were determined by plotting the distribution of *k*-mer coverage weighted with the node length (http://wiki.github.com/dzerbino/velvet/) using the package *plotrix* version 2.9.4 in R version 2.10.1.

6.4.7.5 Contig alignment

Contigs resulting from *de novo* assembly were mapped to the DSM reference genome, using MUMmer version 3.20 (Delcher, Kasif et al. 1999; Delcher, Phillippy et al. 2002; Kurtz, Phillippy et al. 2004). The alignments were done at the nucleotide level using NUCmer with default parameters and the best position for each contig determined with *show-tiling* with a minimum contig alignment coverage of 25. Default parameters were kept for other parameters (minimum percent identity of 90 and minimum contig coverage difference of 10).

6.4.7.6 Visualization

Genbank files were created to visualize the *Streptomyces clavuligerus* DSM reference genome using Artemis version 12 (Rutherford, Parkhill et al. 2000; Berriman and Rutherford 2003). Among the features included in the created Genbank files were contigs, gaps, coding sequences (CDS), and gene products. Mapping results in indexed BAM files were visualized with respect to the DSM reference genome using Bamview within Artemis (Carver, Bohme et al. 2010).

6.5 Results and Discussion

6.5.1 Mutant strain produces more clavulanic acid

Streptomyces clavuligerus WT and *Streptomyces clavuligerus* mutant strains were cultivated in soy-based media in two cultures to characterize their clavulanic acid production. The pH for the cultures was monitored and adjusted to 6.8, as basic conditions degrade clavulanic acid. Growth was measured as optical density at 595 nm (OD_{595}), and did not reveal significant differences between the two strains (Figure 6.2).

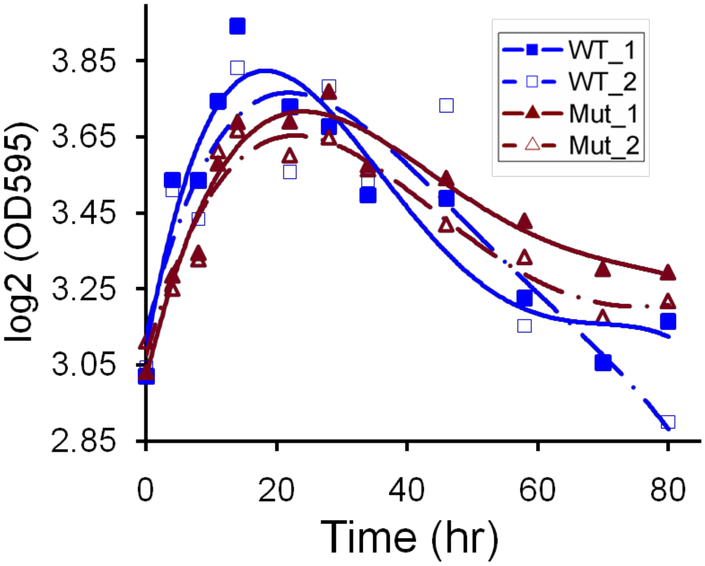


Figure 6.2. Growth for two cultures of each strain. Wild type strain (WT), mutant strain (Mut).

Clavulanic acid production was quantified using high-pressure liquid chromatography (HPLC). A calibration curve was constructed using pure clavulanic acid, a generous gift from Prof. Susan Jensen at the University of Alberta. The linear range of detection was found to be below 150 μg/ml (Figure 6.3).

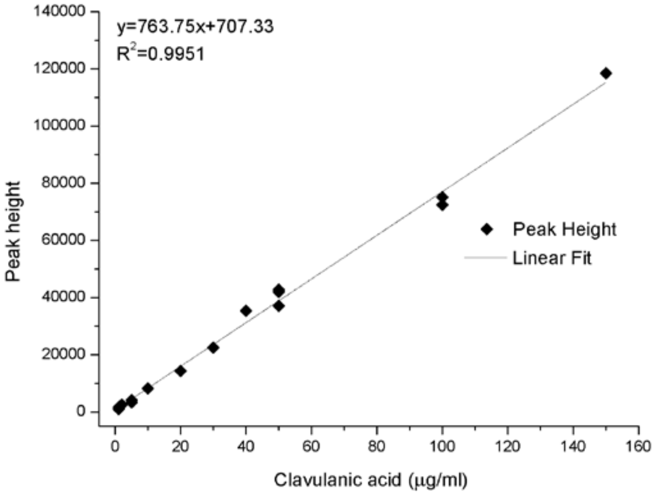


Figure 6.3. HPLC peak height of clavulanic acid standards.

Clavulanic acid levels were found to differ between the strains as early as 22 hours after the start of the fermentation. Maximum levels of clavulanic acid were produced at approximately 33 hours and remained at that level or decreased slightly (due to degradation). In average, the wild type strain produced 110 μg/ml of clavulanic acid, whereas the mutant strain produced 325 μg/ml, a ratio of 3:1. In terms of specific clavulanic acid production (Paradkar and Jensen 1995), the mutant strain showed differences from 33 hours and at the end of the culture the ratio of 3:1 was also obtained.

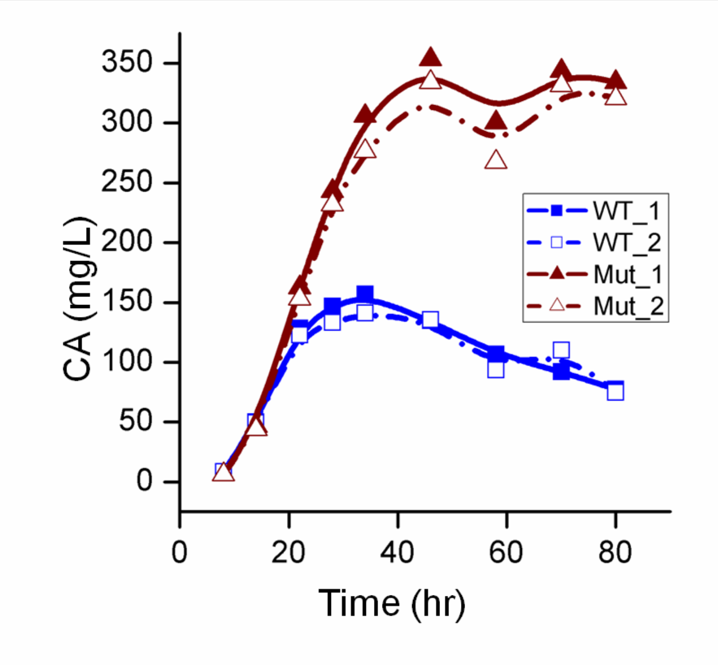


Figure 6.4. Clavulanic acid determination for two cultures of each strain. Wild type strain (WT), mutant strain (Mut).

Table 6.7. Specific clavulanic acid production (μg ml^{-1} OD unit^{-1})

Time (hrs)	WT1	WT2	GSK1	GSK2
8	0.9	0.9	0.7	0.7
14	3.9	4.0	4.2	3.7
33	10.2	10.5	13.3	12.9
28	12.2	10.7	19.7	18.6
34	13.7	12.0	26.0	23.9
46	11.9	11.4	32.6	30.3
58	11.6	9.7	26.2	24.5
70	10.9	12.6	33.1	33.9
80	9.1	9.1	31.1	32.3

6.5.2 Sequencing output

gDNA extracted using the Kirby-mix method, was submitted to NCGR for library preparation and sequencing using Illumina's next generation sequencing technology. The output corresponds to five runs using three paired-end libraries for each strain. This output is summarized in Table 6.8. The received reads were filtered to remove those that contained primer sequences used during library preparation. The resulting clean reads are also summarized in Table 6.8.

6.5.3 Most reads align to a *Streptomyces clavuligerus* reference

The reads obtained from our Solexa sequencing of both strains (WT and mutant) were aligned to the DSM reference using Bowtie. The alignments allowed for up to 3 mismatches, which for these reads (45 and 54 bp) correspond to roughly 93% identity. Although the reads correspond to paired-end libraries, for alignment purposes each side of the paired-end read was treated independently. The results of this mapping can be seen in Table 6.9.

For each run, the percentage of reads mapping to the DSM reference is comparable. A slightly higher percentage of reads from runs 441 and 475 mapped to the DSM reference as compared to runs 517 and 522. All four runs correspond to short insert libraries. However, because of concerns in the quality of the reads towards the end of the read, 4xx runs, which originally had a length of 90 bp, were trimmed to 45 bp by NCGR. The reads from 4xx runs were also filtered, thus it is possible that low quality reads that would otherwise encounter problems in mapping were removed, resulting in a higher percentage of the remaining reads being aligned to the DSM reference. In all cases though, a slightly smaller percentage of reads corresponding to the mutant strain mapped to the DSM reference as compared to the WT reads. This suggested that the mutant strain does indeed have accumulated mutations in its genome. Long insert reads (runs 622 and 625) showed a lower percentage of reads mapped to the DSM reference, and in the case of the mutant strain the difference is noticeable.

Table 6.8. Number of paired-end (PE) reads received, and remaining after filtering.

				WT			Mutant		
Run	Insert type	Library	Length (bp)	Number reads	Number filtered reads	Coverage	Number reads	Number filtered reads	Coverage
441	Short	1	45	2 x 7,226,621	2 x 7,226,022	76	2 x 5,230,096	2 x 5,229,284	55
475	Short	1	45	2 x 5,515,963	2 x 5,515,781	58	2 x 5,697,933	2 x 5,697,553	60
517	Short	2	54	2 x 8,633,540	2 x 8,619,410	109	2 x 14,534,530	2 x 14,323,128	181
522	Short	2	54	2 x 17,647,324	2 x 17,625,482	222	2 x 16,830,926	2 x 16,605,297	210
622/625	Long	3	54	2 x 24,584,052	2 x 24,583,894	310	2 x 22,007,294	2 x 22,001,548	278
TOTAL				**2 x 63,607,500**	**2 x 63,570,589**	**776**	**2 x 64,300,779**	**2 x 63,856,810**	**783**

Table 6.9. Summary of reads mapped to DSM reference.

Run	Strain	Input reads	Reads mapping to chromosome	Reads mapping to plasmid	TOTAL	%
441	WT	14,452,044	10,969,273	2,274,847	13,244,120	92
441	Mutant	10,458,568	8,926,586	533,985	9,460,571	90
475	WT	11,031,562	8,407,666	1,817,327	10,224,993	93
475	Mutant	11,395,106	9,771,843	588,042	10,359,885	91
517	WT	17,238,820	12,168,924	2,442,853	14,611,777	85
517	Mutant	28,646,256	21,567,230	1,899,499	23,466,729	82
522	WT	35,250,964	24,832,706	4,976,550	29,809,256	85
522	Mutant	33,210,594	25,450,840	2,227,465	27,678,305	83
622	WT	24,583,894	32,555,334	6,884,705	39,440,039	80
625	Mutant	22,001,548	30,411,676	1,807,809	32,219,485	73

6.5.4 Low coverage detected in the extremes of the mutant strain plasmid

From the mapping results, coverage was calculated for each strain and genomic element (Table 6.10). For both strains the coverage for the plasmid was lower than the coverage for the chromosome. In the WT strain, the coverage of the plasmid represents 77% of that in the chromosome. Although this difference could be a result of bias in gDNA isolation or library preparation, it is unlikely. The plasmid has a linear structure and considering its size (1.8 Mb) would behave more like a chromosome in terms of gDNA extraction. This difference in coverage could also indicate that some of the cells in the population sequence have actually lost the plasmid. Survival without the plasmid seems possible, as the plasmid does not contain any essential genes of primary metabolism (Medema, Trefzer et al. 2010). In the case of the mutant however, the coverage in the two replicons is quite different. The coverage in the plasmid represents only 28% of that of the chromosome. This difference is large enough to be considered bias resulting from gDNA isolation, library, or sequencing. Both replicons have a G+C content of 72%.

Table 6.10. Coverage calculated from aligned reads.

Strain	Chromosome	Plasmid	Coverage ratio (Plasmid/Chromome)
WT	685	532	0.77
Mutant	743	206	0.28

Visualization of the mapped reads using BamView within Artemis revealed that while coverage presents peaks and valleys, influenced by G+C content, two sharp changes appear in the case of the plasmid for the mutant strain. The edges of the plasmid have extremely low coverage, when compared to the central region of the plasmid. This “boundaries” occur on the 5’ end within locus SCLAV_p0761, and on the 3’end between loci SCLAV_p1267 and SCLAV_p1268. These boundaries are shown in Figure 6.5, in which the DSM reference is represented with the gray lines (forward and reverse strand), and coding sequences are represented by light blue arrows. The reads from our sequencing project are represented on top as green horizontal lines when the read is duplicated and as dark gray horizontal lines when it is unique.

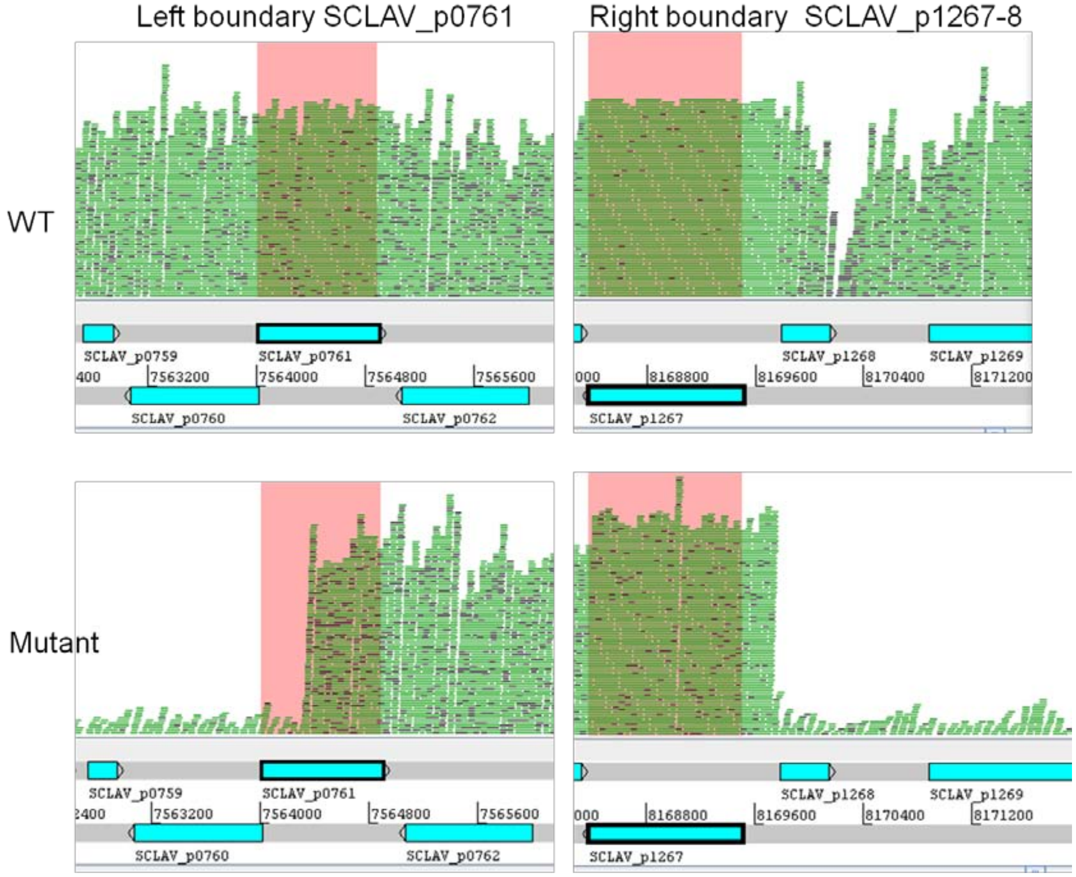


Figure 6.5. Coverage differences detected in mutant strain. Top panels: WT strain. Bottom panels: Mutant strain. RHS panels: 3'end, between genes SCLAV_p1267 (highlighted) and SCLAV_p1268. LHS: 5'end, located within gene SCLAV_p0761 (highlighted).

Coverage of the plasmid in the mutant strain appears as low in the 5'end in a region that extends for 804 Kb, with an exception of a small "island" located from 420-435 Kb. In this "island" the coverage level is similar to that in the central region of the plasmid. The central region, with regular coverage, extends for 606 Kb. Finally, the 3'end of the plasmid, a region extending 387 Kb, presents low coverage (Figure 6.6).

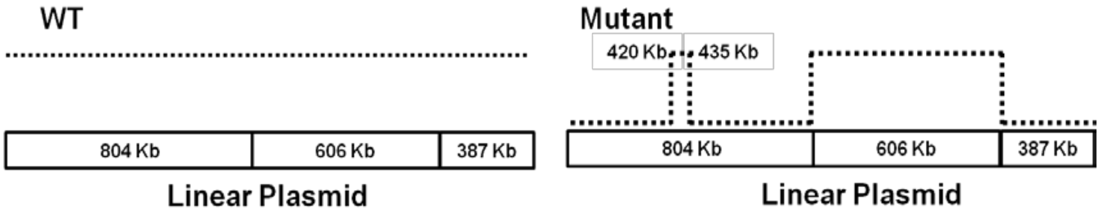


Figure 6.6. Schematic representation of coverage in the linear plasmid. The mutant strain has low coverage in both ends of the plasmid, and regular coverage in its central region. An "island" of regular coverage was detected in the left region of the plasmid.

Streptomyces chromosomes are large linear structures, consisting of a “core” in which most of the primary metabolism genes are located, and two “arms” in which most of the clusters for secondary metabolite clusters are located. The small size of the chromosome in the DSM reference and the high number of secondary metabolite clusters located in the plasmid lead Medema et al. to investigate the possible origin of the megaplasmid. Comparison to other *Streptomyces* genomes revealed that the chromosome corresponds to the core of other *Streptomyces* genomes, and that the “arms” are missing. The extremes of the megaplasmid are similar to the “arms” regions of other *Streptomyces* chromosomes, thus they suggest that either a double crossover or two subsequent recombinations between an originally smaller plasmid and the “arms” of the chromosome occurred in the recent past.

The difference in coverage detected in the mutant strain in our project suggests that in most of the cells in the sequenced population, the extremes of the linear plasmid have been lost. When the coverage for the plasmid is recalculated for a 621 Kb size (606 Kb central region + 15 Kb island), it changes from the original 206x (Table 6.10) to 597x, thus making the ratio $\text{Coverage}_{\text{plasmid}}/\text{Coverage}_{\text{chromosome}} = 0.80$, a value very similar to that calculated for the WT strain (0.77).

The missing regions could correspond to the “arms” with chromosomal origin proposed by Medema et al. The genes that originally alerted Medema et al. of a possible chromosomal origin for parts of the current megaplasmid include *parAB*, *tap,* and *tpg*, all involved in replication. These genes, integrally related to the plasmid survival are within the 606 Kb central region with regular coverage. Thus a short plasmid of approximately 621 Kb would be capable of replication and fully functional.

6.5.5 qRT-PCR confirmation of deleted regions in the plasmid of the mutant strain

Quantitative Real Time Polymerase Chain Reaction (qRT-PCR) was used to confirm the difference in coverage level. Six genes were chosen for this confirmation. A gene was chosen in the 5’ end of the plasmid, a region with low coverage. A gene in the “island” of regular coverage (region 420-435 Kb) was also included in this confirmation. Genes located to just before and after the coverage boundaries were also included in this analysis. A schematic representation of the position of the selected genes can be seen in Figure 6.7. Details of the gene tested by qRT-PCR can be found in Table 6.11.

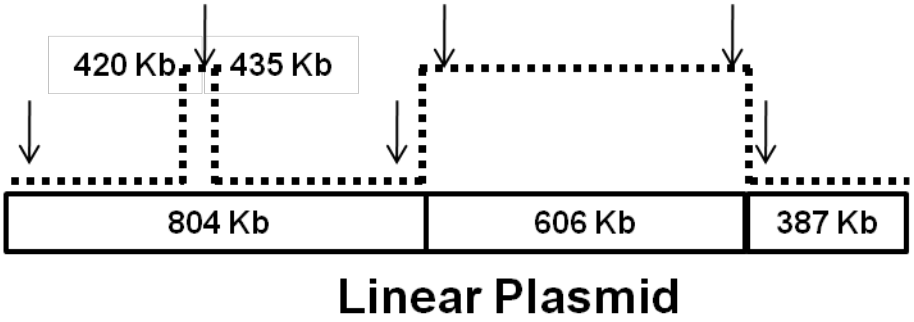


Figure 6.7. Schematic representation of the location of genes tested with qRT-PCR, indicated by arrows.

Table 6.11. Genes selected for testing by qRT-PCR

Locus	Reason for testing	gDNA amount used (ng)
SCLAV_p0171	Left region (low coverage)	5
SCLAV_p0440	Island in left region (regular coverage)	5
SCLAV_p0760	Before left boundary (low coverage)	25
SCLAV_p0762	After left boundary (regular coverage)	5
SCLAV_p1266	Before right boundary (regular coverage)	5
SCLAV_p1269	After right boundary (low coverage)	25

Results from qRT-PCR are presented as ∆Ct, calculated as Ct(Gene of interest) – Ct(*gyrA*). Using this calculation, positive numbers indicate low gene levels, corresponding to a higher number of PCR cycles before the gene was detected. Whereas most of the genes tested have a ∆Ct value within ± 2 for the WT strain (the exception being SCLAV_p1269 with ∆Ct = 3.72), a clear difference is seen for the genes in the mutant strain, according to their location. The genes in the regular area of coverage (SCLAV_p0440, SCLAV_p0762, and SCLAV_01266) present ∆Ct values similar to those of the WT strain (within ± 2). But the genes in the low coverage area (SCLAV_p0171, SCLAV_p0760) have extremely high ∆Ct values (> 12), indicating the low level of the gene in the genome. Notice that for SCLAV_p0760 five times more template gDNA was used. Furthermore, SCLAV_p1269 could not be detected even when the gDNA template was increased to 25 ng.

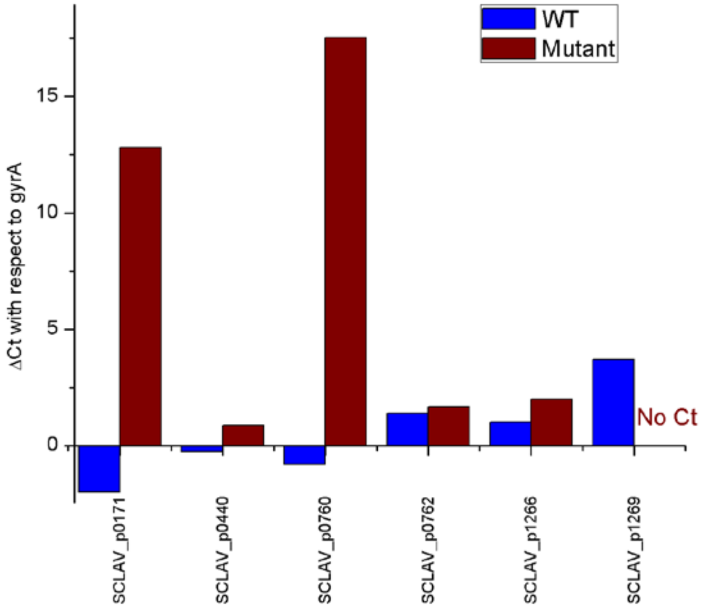


Figure 6.8. qRT-PCR results expressed as ΔCt with respect to *gyrA*. Higher values indicate lower levels in gDNA.

6.5.6 Genome-wide variant analysis

The reads mapped to the DSM reference were used to determine single nucleotide changes (variants) in the two strains using SAMtools. Even though the WT strain used in our sequence, *Streptomyces clavuligerus* NRRL 3585, corresponds to the same isolate as that used in the work by Medema et al., *Streptomyces clavuligerus* ATCC 27064 (Higgens and Kastner 1971), reads from both WT and mutant strains were analyzed for variants. This approach accounts for possible rates of spontaneous mutations. In addition differences between the two WT strains are also of interest, as they could represent true differences or sequencing errors.

Table 6.12 summarizes the number of variants detected in each strain and genomic element with respect to the DSM reference. It is interesting to note that in the case of the WT strain, the number of variants detected for the chromosome and the plasmid are very similar, however considering the size of each, the mutation rate is higher in the plasmid (1 per 2,747 nucleotides) than in the chromosome (1 per 10,060 nucleotides). The mutation rate for the mutant strain is very similar to that of the WT for the chromosome (1 per 9,223 nucleotides), but it is extremely high for the plasmid (1 per 418 nucleotides).

Table 6.12. Summary of single nucleotide changes found in each strain and genomic element.

Strain	Chromosome	Plasmid
WT	672	654
Mutant	733	4,297

Changes in single nucleotides, however, can be supported by only a portion of the reads covering that particular position. It was thus of interest to identify the distribution of this support to determine a threshold above which there was certainty of the change. As can be seen in Figure 6.9, for both strains and genomic elements, the percentage of reads supporting the change peaks at roughly 20%. However, in all four cases the distribution extends to a percentage of reads of up to 91%, at which point the distribution is cut, and reappears at 100%. The term mutation will be used to refer to single nucleotide changes supported by 100% of the reads, whereas those with lower support will be called variants. Considering that the Solexa sequencing error is in the order of 0.05-0.12% (Cronn, Liston et al. 2008; Rougemont, Amzallag et al. 2008) it is then safe to assume that single nucleotide changes supported by 100% of the reads indeed represent mutations.

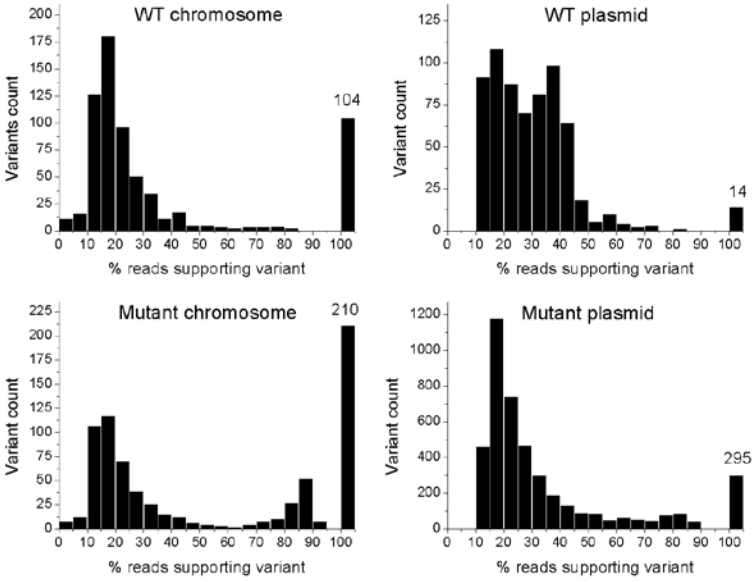


Figure 6.9. Percentage of reads supporting base change.

The focus in this study is in those mutations that appear only in the mutant strain, thus those mutations or variants appearing in the WT strain also were removed from further analysis. Only 117 mutations in the chromosome passed these criteria, and 51 in the plasmid.

The open reading frames (ORFs) predicted in the work by Medema et al. were used to determine the number of mutations located within ORFs. Mutations within ORFs are of higher interest when the change is non silent, that is, the translated amino acid sequence is changed. The mutant strain contains 58 non silent mutations in the chromosome and 27 in the plasmid (Figure 6.10).

All intergenic mutations are of interest, as that could affect the binding of regulatory proteins, or of RNA polymerase, but the change can affect only coding sequences located downstream from it. Intergenic mutations were classified as appearing downstream of two ORFs, upstream of one ORF, or upstream of two ORFs. The mutations of interest in this work are those that appear upstream of one or two ORFs (Figure 6.10). This schema for filtering is shown in Figure 6.10, where mutations of high interest appear in blue.

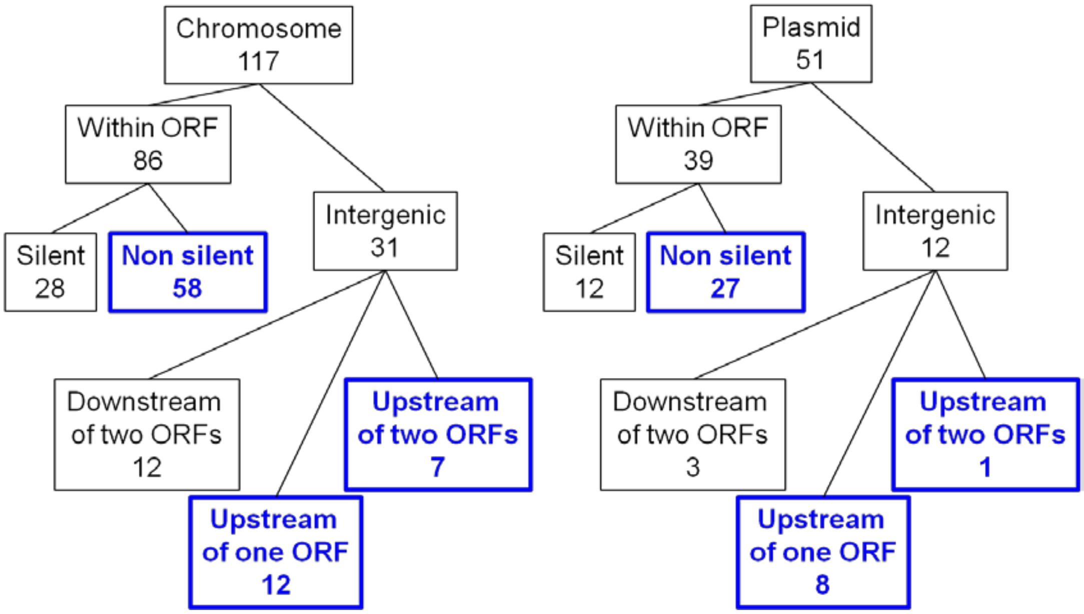


Figure 6.10. Mutations detected in mutant strain only. The mutations within ORFs are of high interest when they result in non silent changes. Intergenic mutations are of interest when they are upstream of at least one coding sequence. High interest mutations appear in blue in this schema. The number of mutations detected in each case is also given.

6.5.7 Key mutations detected in the mutant strain

The complete list of 85 (58 in the chromosome + 27 in the plasmid) non silent mutations can be found in Appendix B, Tables B.1 and B.2. Of the 85 mutations, ten occurred within genes belonging to secondary metabolite clusters. Of these high interest mutations four occur in the chromosome and six in the plasmid (Table 6.13).

Two of the mutations occur in genes which are part of the macrolide type I PKS. This cluster (the only one for a macrolide) was proposed by Medema et al. to be the biosynthetic cluster for the tacrolimus-like macrolide previously reported by (Kim and Park 2008).

Table 6.13. Non silent mutations in coding sequences belonging to secondary metabolite clusters.

Position	Reads	Locus	Annotation	Cluster
27770	806	SCLAV_0012	Modular polyketide synthase	Macrolide type I PKS
63639	532	SCLAV_0015	Modular polyketide synthase	Macrolide type I PKS
4885031	805	SCLAV_4205	3'-hydroxymethylcephem-O-carbamoyltransferase	NRPS/Beta-lactam, cephamycin C
6211944	873	SCLAV_5271	Non-ribosomal peptide synthetase	NRPS
752810	7	SCLAV_p0711	Condensation domain protein	Other
1297373	371	SCLAV_p1172	Undecaprenyl pyrophosphate synthetase	Terpene synthase, pentalenene synthase (2x)
1440512	16	SCLAV_p1283	Moenomycin biosynthesis protein MoeGT4	Phosphoglycolipid, moenomycin gene cluster
1544237	8	SCLAV_p1355	Hypothetical protein	Enediyne PKS
1558250	31	SCLAV_p1367	Peptidase	Enediyne PKS
1572072	5	SCLAV_p1376	Subtilisin-like protease	Enediyne PKS

The next non silent mutation within a gene part of a secondary metabolite cluster was in the *cmcH* gene (SCLAV_4205). *cmcH* encodes a 3'-hydroxymethylcephem-O-carbamoyltransferase, part of the cephamycin C biosynthetic cluster. The change C → occurs at position 163. This change would result in an amino acid change from proline, a non-polar amino acid, to serines, a polar mino acid with positive charge. If this change renders the enzyme inactive, carbamoylation of deacetylcephalosporin-C (DAC) would not occur. DAC would then accumulate in the media and cephamycin C would not be produced (Figure 6.12).

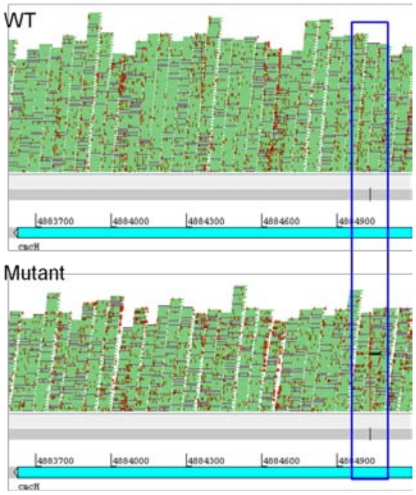


Figure 6.11. Non silent mutation within the cmcH gene in the mutant strain. The mutation is seen as a vertical red line (blue box).

The last non silent mutation within a secondary metabolites cluster in the chromosome occurs in SCLAV_5271. This gene is part of one of the multiple nonribosomal peptide synthetase (NRPS) clusters present in the *Streptomyces clavuligerus* genome.

The first two non silent mutations in secondary metabolite clusters in the linear plasmid occur within clusters for which no additional information is available. The next non silent mutation detected occurs in the glycosyl transferase MoeGT4, located in the plasmid and part of the moenomycin biosynthetic cluster. Moenomycin is a phosphoglycolipid antibiotic which inhibits peptidoglycan biosynthesis and that has potential as the starting point for novel antibiotics (Makitrynskyy, Rebets et al. 2010). MoeGT4 is an early enzyme in the moenomycin pathway, attaching a sugar to form a disaccharide (Ostash, Doud et al. 2009).

The final three non silent mutations occur within genes part of the enediyne secondary metabolite cluster, but none of them correspond to the *unb* genes, which are responsible for the core structure of enediyne.

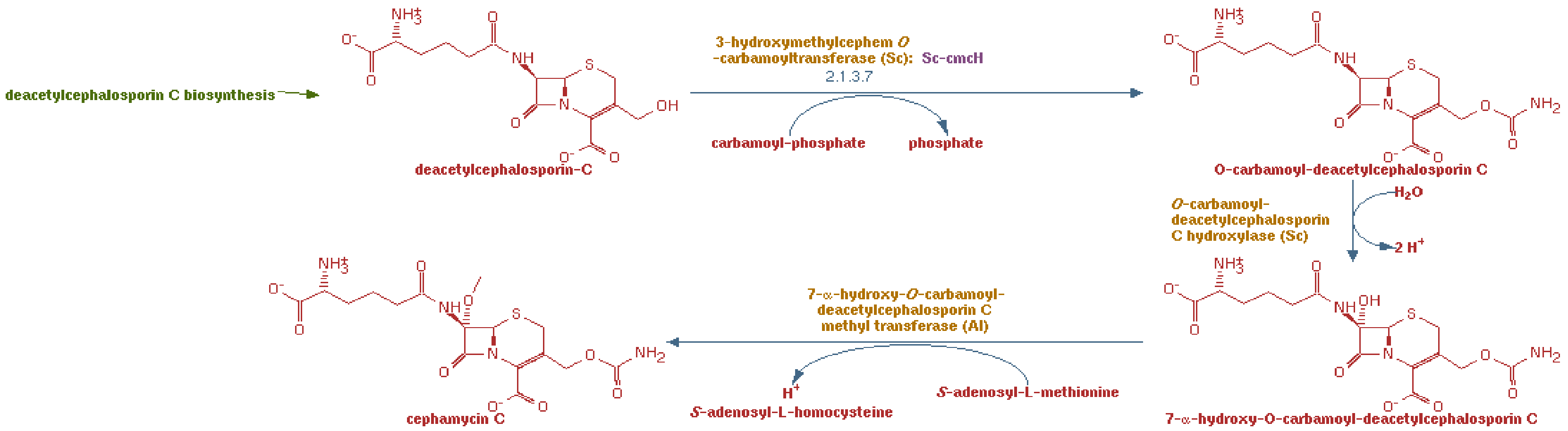

deacetylcephalosporin C biosynthesis
deacetylcephalosporin-C
3-hydroxymethylcephem *O* -carbamoyltransferase (Sc): Sc-cmcH
2.1.3.7
carbamoyl-phosphate
phosphate
O-carbamoyl-deacetylcephalosporin C
O-carbamoyl-deacetylcephalosporin C hydroxylase (Sc)
H_2O
2 H^+
7-α-hydroxy-*O*-carbamoyl-deacetylcephalosporin C methyl transferase (Al)
S-adenosyl-L-methionine
H^+
S-adenosyl-L-homocysteine
cephamycin C
7-α-hydroxy-O-carbamoyl-deacetylcephalosporin C

Figure 6.12. Cephamycin C biosynthesis (*cmcH* appears in purple, Sc: *Streptomyces clavuligerus*). Image from Metacyc (http://biocyc.org/).

Of the non silent mutations that are not part of a secondary metabolite cluster (Appendix B Tables B.1 and B.2), three are of particular interest: SCLAV_2321, SCLAV_3643, and SCLAV_4545. SCLAV_2321 encodes a penicillin acylase. Penicillin acylases are of industrial importance, as they are used to cleave the side chain of penicillins, and a new side chain is then attached to produce a variety of penicillins. The *in vivo* role of penicillin acylases, however, remains unclear (Duggleby, Tolley et al. 1995). SCLAV3643 (*rpsC*) encodes a 30S ribosomal protein S3, and SCLAV_4545 (*rplS*) a 50S ribosomal protein L19.

A total of 20 (Appendix B Tables B.3 and B.4) single nucleotide mutations were found in the intergenic regions upstream of a coding sequence. This type of mutation is relevant since it could change the strength of the promoter region, and thus change the expression level of the downstream coding sequence. Table 6.1 includes the intergenic mutations that are upstream of a gene which is part of a secondary metabolite cluster. Two of them are contiguous (1,307,226-1,307,227 in the plasmid), upstream of the thioesterase SCLAV_p1183 in the pentalenene synthase (2x) cluster.

Table 6.14. Intergenic mutations which are upstream of one coding region belonging to a secondary metabolite cluster.

Position	Reads	Downstream locus	Annotation	Cluster
4882915	323	SCLAV_4204	Positive regulator	NRPS/Beta-lactam, cephamycin C
1307226	46	SCLAV_p1183	Thioesterase	Terpene synthase, pentalenene synthase (2x)
1307227	39	SCLAV_p1183	Thioesterase	Terpene synthase, pentalenene synthase (2x)

At position 4,882,915 in the chromosome a G → A mutation was detected in the mutant strain. This mutation was supported by 100% of the 323 reads mapping to the position (Figure 6.13). This mutation is upstream of the positive regulator *ccaR* (SCLAV_4204). *ccaR* is located in the cephamycin C cluster (Figure 6.14) and is required for the production of both antibiotics (Perez-Llarena, Liras et al. 1997). CcaR belongs to the *Streptomyces* antibiotic regulatory proteins (SARP) family and it was found to bind to its own promoter region (Santamarta, Rodriguez-Garcia et al. 2002).

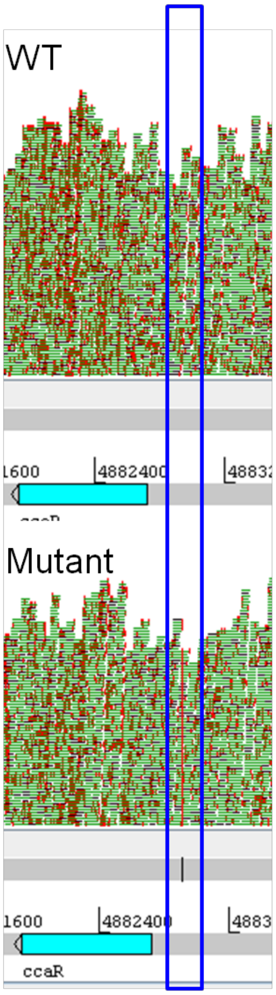


Figure 6.13. Intergenic mutation upstream of the regulator *ccaR*. The mutation can be seen as a red vertical line in the case of the mutant strain.

The *ccaR* translation coding sequence starts at position 4,882,712, thus the mutation is located at -203 bp. Two transcription start points are known for *ccaR*, the first at -74 bp and the second at -173 bp (Wang, Tahlan et al. 2004). The -35 promoter region identified by Wang et al. covers the range -206 to -211 bp, putting the mutation found in our study only two nucleotides away from it and most likely responsible for the high clavulanic acid production phenotype in the mutant strain. Furthermore, (Santamarta, Rodriguez-Garcia et al. 2002) reported the binding of CcaR to its own promoter region, and although the exact position of its binding was not determined, it is known to bind a region that extends up to position -234. This mutation, could then affect *ccaR*'s -35 promoter region or the region bound by CcaR, or both.

This mutation does not occur within the AutoRegulatory Element (ARE) box described by (Santamarta, Perez-Redondo et al. 2005). ARE boxes have been found upstream of genes encoding SARP proteins. ARE boxes have been found to be bound by butyrolactone receptor proteins. In the case of *Streptomyces clavuligerus*, the ARE box upstream of *ccaR* ($\text{ARE}_{\textit{ccaR}}$) is located at -890 bp.

A mutation in an intergenic region can be upstream of two coding sequences. Eight mutations (Table 6.15) of this type were detected in total, seven in the chromosome and 1 in the plasmid. In two of the eight cases, positions 1,043,217 and 2,794,732, bp of the chromosome, only a few reads mapped the position (3 and 6 respectively).

The mutation at position 1,856,565 bp in the chromosome is upstream of SCLAV_1555 and SCLAV_1556. This mutation is of interest as SCLAV_1556 encodes a β-lactamse. The mutation is located at -28 bp from SCLAV_1555 and at -76 bp from SCLAV_1556. These distances are well within a typical size for a prokaryotic promoter, thus further studies are required to test if this mutation affect the transcription of either gene.

Two continuous mutations (1,884,243 – 1,884,244) are located in the intergenic region between SCLAV_1579 and SCLAV_1580. The mutations are at -98 and -99 bp from the hypothetical protein SCLAV_1579, and at -31 and -32 bp of the new domain nuclease-related domain (NERD domain)-containing protein. Proteins containing the NERD domain were suggested to have a nuclease function (Grynberg and Godzik 2004).

The next mutation upstream of two coding sequences is located at 2,766,473 bp in the chromosome. The mutation is at approximately the same distance from the translation start of the two adjacent coding sequences. The mutation is located at position -288 of *hrdD* and at -262 of the hypothetical protein SCLAV_2336. HrdD is not essential in *Streptomyces coelicolor* (Buttner, Chater et al. 1990; Buttner and Lewis 1992), but was found to be expressed in *Streptomyces coelicolor*, *Streptomyces aereofaciens*, and *Streptomyces griseus*. In *Streptomyces aureofaciens* and *Streptomyces griseus*, *hrdD* is expressed only in certain stages of growth. In *Streptomyces aureofaciens*, the *hrdD* transcript was detected during the vegetative stage (Kormanec, Farkasovsky et al. 1992; Kormanec and Farkasovsky 1993). In *Streptomyces griseus* it was detected during sporulation and during growth in phosphate-rich medium (Marcos, Gutierrez et al. 1995).

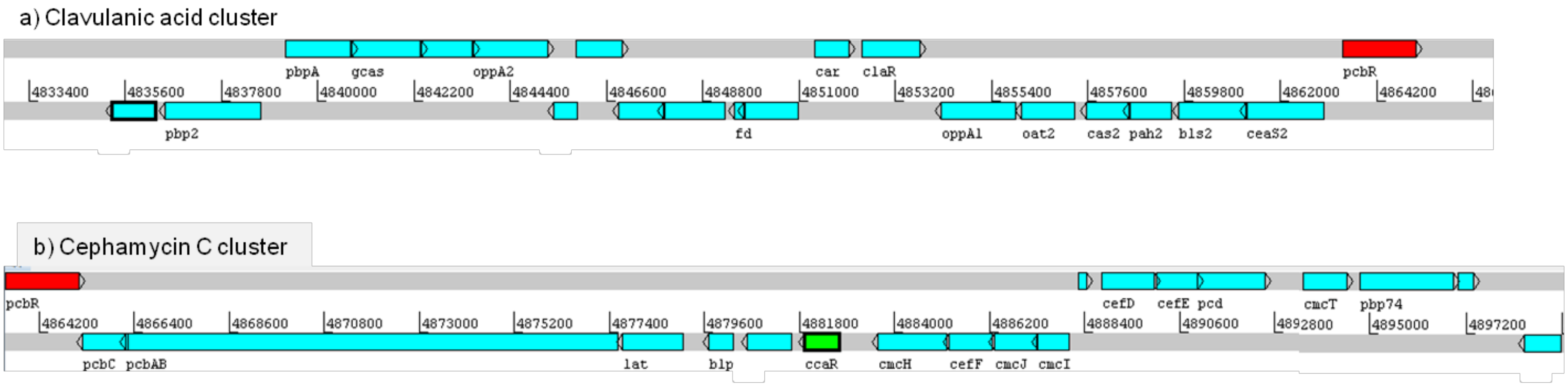


Figure 6.14. a) Clavulanic acid cluster, b) Cephamycin C cluster. *pcbR* (in red) is the first gene in the cephamycin C cluster. Notice its position next to the clavulanic acid cluster. The positive regulator *ccaR* (in green) is known to affect product of both beta-lactam antibiotics. A mutation was found in the upstream region of *ccaR* in the mutant strain.

For the mutation at position 3,541,555 bp in the chromosome, even though it is upstream of two coding sequences the mutation is not likely to have an effect on the M18 family aminopeptidase SCLAV_3006, as the mutation is located at -672 bp from its translational start point. The mutation closer to the translation start point for SCLAV_3007, which encodes a putative DNA-binding protein.

The mutation detected in the plasmid is in the region of extremely low coverage in the mutant strain, which is suspected to be a deletion, and thus will not be discussed.

Table 6.15. Mutations upstream of two coding sequences. None of the coding sequences is part of a secondary metabolite cluster.

Position	Reads	Locus_one	Annotation	Locus_two	Annotation
1043217	3	SCLAV_0830	Hypothetical protein	SCLAV_0831	Predicted amino acid aldolase or racemase
1856565	874	SCLAV_1555	Aldehyde dehydrogenase	SCLAV_1556	Beta lactamase
1884243	947	SCLAV_1579	Hypothetical protein	SCLAV_1580	NERD domain-containing protein
1884244	937	SCLAV_1579	Hypothetical protein	SCLAV_1580	NERD domain-containing protein
2766473	673	SCLAV_2335	RNA polymerase principal sigma factor hrdD	SCLAV_2336	Hypothetical protein
2794732	6	SCLAV_2357	Hydrolase	SCLAV_2358	Tryptophanyl-tRNA synthetase 1
3541555	919	SCLAV_3006	M18 family aminopeptidase	SCLAV_3007	Putative DNA-binding protein
10262	55	SCLAV_p0009	Putative TraA protein	SCLAV_p0010	40-residue YVTN family beta-propeller repeat protein

6.5.8 Effect of a mutation detected upstream of the regulator CcaR

The presence of a mutation upstream of the coding sequence of the regulator *ccaR* suggests the possibility of a change in transcript level for this gene. If a difference in the protein level also occurs, this could have an effect on the production of clavulanic acid and cephamycin C. In an ongoing experiment, we measured the transcript level of *ccaR* by qRT-PCR in culture samples collected at different time points for both strains. Other genes included in this measurement were the known CcaR targets *cefD* and *cmcI*. As no direct CcaR targets are known in the clavulanic acid pathway, the transcript level of early, middle, and late step enzymes was also measured. The list of genes included in the measurement is given in Table 6.16. The transcript levels were measured at 16 and 32 hours.

Table 6.16. Genes included in transcript level measurement by qRT-PCR

Gene	**Role**
ccaR	Regulator
claR	Regulator
car	Late step enzyme
cas2	Middle step enzyme
gcas	Middle step enzyme
ceaS2	First step enzyme
cefD	ccaR target
cmcI	ccaR target
cmcH	Cephamycin, non silent mutation

The results are presented as $-\Delta$Ct (Ct_{gyrA}-Ct_{goi}) in Figure 6.15, thus the higher the value, the higher the expression level. With the exception of *claR* in the WT, the rest of the transcripts had a higher expression level in the second time point, 32 hours. In the mutant strain, however, the difference in transcript level is more dramatic. The known CcaR target *cefD* shows the biggest difference in transcript level between the two time points. The *cmcH* transcript also presents a high difference in transcript level between the two time points.

In the case of clavulanic acid related transcripts, the biggest difference occurs for the *ccaR* regulator itself, followed by the middle step enzymes *cas2* and *gcas*. Although preliminary, these results point to possible CcaR targets in the clavulanic acid cluster. As the rest of the time points and cultures are tested, it will be interesting to see if the middle step enzymes are the most affected transcripts in this clavulanic acid high producer strain.

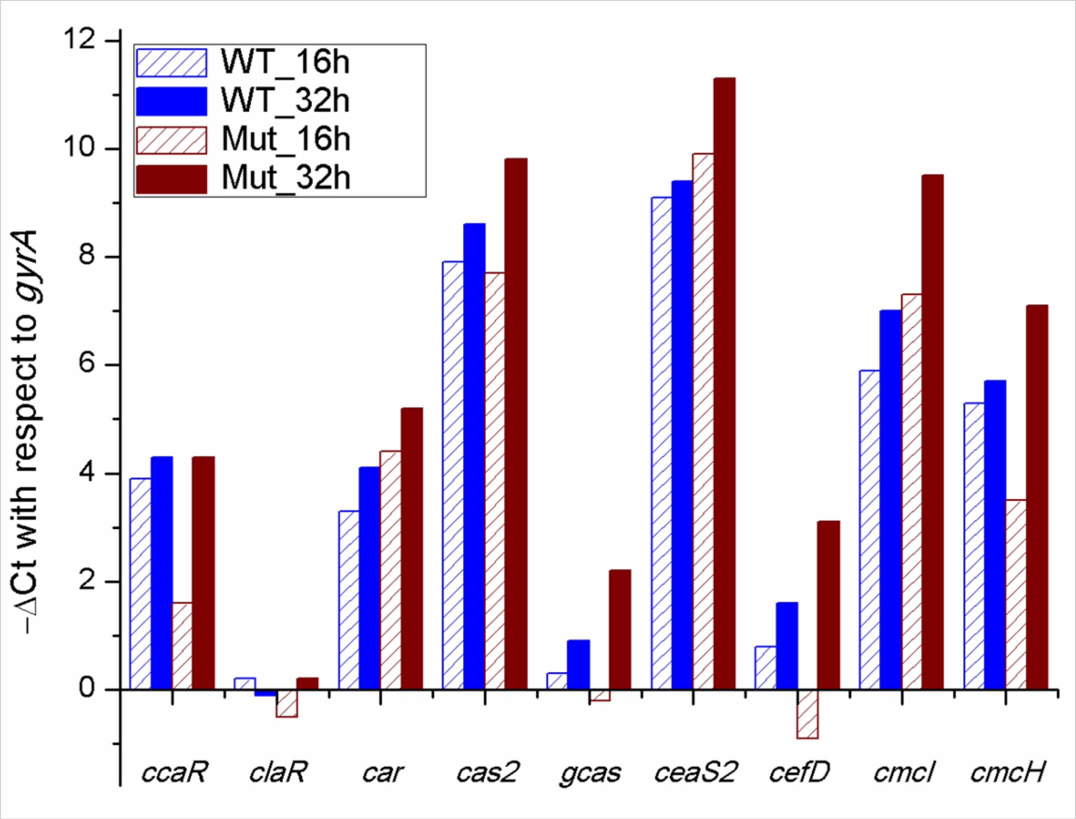


Figure 6.15. qRT-PCR results for transcript level measurements.

6.5.9 Other differences detected between the strains

A total of 113 mutations were detected in both strains when compared to the DSM reference, 13 in the plasmid (Appendix B, Table B.5) and 90 in the chromosome (Appendix B, Table B.6). Since the mutation appears in both, the WT and mutant strains sequenced in this work, and the nucleotide change is the same in both strains, these differences could hardly be explained as true differences, but rather sequencing error.

Fourteen mutations were detected when reads from our WT strain were mapped to the DSM reference (Table 6.17). These changes either do not appear in the mutant strain or are not supported by 100% of the reads. Since the WT strains used in both projects correspond to the same isolate, these differences suggest either sequencing errors in either project or true differences arising during the lab life of the strains.

However, it is interesting to note that the number of reads mapping those positions is low, with the maximum number being twelve. This could then suggest a threshold for mutation confidence.

Table 6.17. Mutations detected when the WT reads were mapped to the DSM reference.

Position	Base in DSM reference	Base in WT	Reads	Locus
860938	T	C	4	Intergenic
860939	A	C	6	Intergenic
2576364	T	C	12	Intergenic
2576365	T	C	9	Intergenic
2592090	A	T	6	Intergenic
2609349	A	C	3	Intergenic
3243682	T	C	3	Intergenic
4384880	C	G	4	Intergenic
4577733	C	G	4	Intergenic
4577734	T	G	11	Intergenic
4926475	A	G	7	Intergenic
1043106	C	G	10	SCLAV_0830
1043108	G	C	7	SCLAV_0830
1276747	T	C	6	Intergenic

6.5.10 *De novo* assemblies

Two types of assemblies were performed for each strain using Velvet:

1. Reads from all three libraries were treated as unpaired and combined in a single file for input into the assembler (unpaired assembly).
2. Reads from the three libraries were kept separate, and the information on pairing was used as input into the assembler. This type of assembly requires an estimate of the insert size and standard deviation. These statistics (Table 6.18) were obtained from mapping paired end reads to the DSM reference and using Artemis for their visualization.

Table 6.18. Insert length and standard deviation for each of the three libraries for each strain.

	WT		Mutant	
Library	Length (bp)	Std. dev. (bp)	Length (bp)	Std. dev. (bp)
1 (4xx)	425	25	450	50
2 (5xx)	325	25	375	25
3 (62x)	3500	500	3250	450

The expected coverage and coverage cutoff parameters were optimized to obtain the final assemblies. Statistics for the unpaired assemblies appear in Table 6.19.

Table 6.19. Unpaired assembly results.

Case	Input reads	Number contigs	Min. contig length (bp)	Max. contig length (bp)	N50 (bp)	Reads assembled	% reads assembled	Total length (Mb)
WT	127,141,178	1,334	100	101,850	23,476	117,657,202	93	9.0
Mutant	127,713,620	1,143	100	144,115	24,351	117,315,929	92	7.9

From the statistics in Table 6.19 it is clear that assemblies are comparable. A very similar percentage of reads was successfully assembled into contigs, the number of contigs and the N50 for them are in the same order of magnitude. An important difference though is the total length of the genome, calculated by adding the length of all the contigs. The assembled genome for the mutant strain is 7.9 Mb, whereas that for the WT is 9.0 Mb. Thus the assembled genome for the mutant strain is smaller than that for the WT strain, and slightly smaller than other *Streptomyces* genomes. The assembled genome for the WT strain however is slightly larger than that published for the DSM reference (8.6 Mb). This could indicate that there might still be some overlap between the contigs, but that the assembler failed to join.

Table 6.20 has the results for the second case, paired assemblies. Even though the percentage of reads assembled and the total length of the genome are practically the same as for the case of unpaired assemblies, there is a great difference in the contig statistics. The number of contigs has reduced considerably (from more than 1000 to less than 650) and the maximum contig length and N50 have increased considerably.

Although scaffolding information was obtained for the paired assemblies it will not be presented in detail, as the resulting total length (11.4 Mb for the WT and 9.6 Mb for the mutant) appears considerably longer than the expected genome sizes. De novo assemblies were also constructed with SOAPdenovo, however the results were poor even when compared with those obtained by Velvet without optimization.

Table 6.20. Paired assembly results.

Strain	Input reads	No. of elements	Min (bp)	Max (bp)	N50 (bp)	Reads assembled	% reads assembled	Total length
WT	127,141,178	460	100	314,855	67,964	117,829,150	93	9
Mutant	127,713,620	628	100	230,404	45,627	117,461,776	92	8

6.5.11 Contig tiling confirms that the mutant strain has lost the plasmid extremes.

The assembled contigs were aligned to the DSM reference using NUCmer, and the best possible position for each contig determined using *show-tiling*. NUCmer and show-tiling are part of MUMmer, a software for the rapid alignment of large DNA (or amino acid) sequences. Contigs with a minimum contig alignment coverage of 25 were selected for output. This setting was selected as the minimum length for the assembled contigs is 100 bp, thus at least 25 bp are required to map to the DSM reference. As the DSM reference is not fully completed, and it still contains gaps (stretches of Ns), no alignment is valid to those regions. The number and percentage of contigs that were successfully tiled using the above mentioned parameters are summarized in Table 6.21.

Table 6.21. Number of contigs tiled to the DSM reference. The percentage with respect to the total number of contigs appears in parenthesis.

Strain	Unpaired assembly	Paired assembly
WT	1009 (76%)	248 (54%)
Mutant	915 (80%)	459 (73%)

For the case of contigs generated by the unpaired assembly, the percentage of contigs successfully tiled to the DSM reference is very similar. But for the case of the paired assembly though, the percentage of contigs from the WT strain that successfully tile is almost 20% less than that of the mutant (Table 6.21). However, when the tiling results are analyzed in the context of total length, the results are quite different (compare Table 6.21 and Table 6.22).

Table 6.22. Length (Kb) mapped by the assembled contigs. For the length calculation the overlapping areas were subtracted. The percentage of the DSM length mapped appears in parenthesis.

	Unpaired		Paired	
Strain	Chromosome	Plasmid	Chromosome	Plasmid
WT	6610 (98%)	1719 (96%)	6552 (97%)	1668 (93%)
Mutant	6605 (98%)	662 (37%)	6601 (98%)	641 (36%)

When analyzed in terms of length, it is clear that even when the percentage of WT contigs mapping the DSM reference is lower, in terms of length the mapping is almost the same for the chromosome and actually much larger for the plasmid. These results are in agreement with the direct mapping of reads to the DSM reference (Section 6.5.4), in which a drastic difference is coverage was observed.

This alignment confirmed that while the contigs from the WT strain cover the DSM reference almost completely, the contigs from the mutant strain leave the edges of the DSM reference's plasmid unaligned. The total aligned length calculated here is in the same order of magnitude as that calculated from read mapping (662 Kb for unpaired, 641 Kb for paired, and 606 Kb for read mapping). The larger length for the mapping of the contigs generated from the unpaired assembly can be explained by the many contigs that do map the regions 0-804 Kb and 1410-1797 Kb, which correspond to the suspected deleted regions. The contigs that do map this region, however, are short (215 bp in average) and are separated by large gaps (5400 bp in average). The alignment of assembled contigs to the DSM reference can be seen in Figure 6.16.

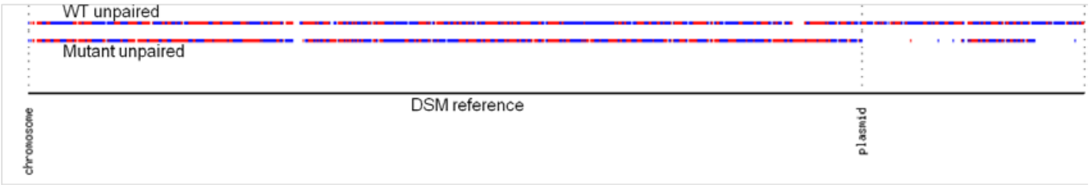


Figure 6.16. Contigs generated by de novo assembly aligned to the DSM reference. The DSM reference is indicated by the black line. Red rectangles: contigs mapping to reference in the direction +/+. Blue rectangles: contigs mapping to reference in the direction -/+.

Although including the information on paired-reads resulted in improved assemblies in terms of fewer and longer contigs, the mapping obtained from these contigs is more fragmented (Figure 6.17).

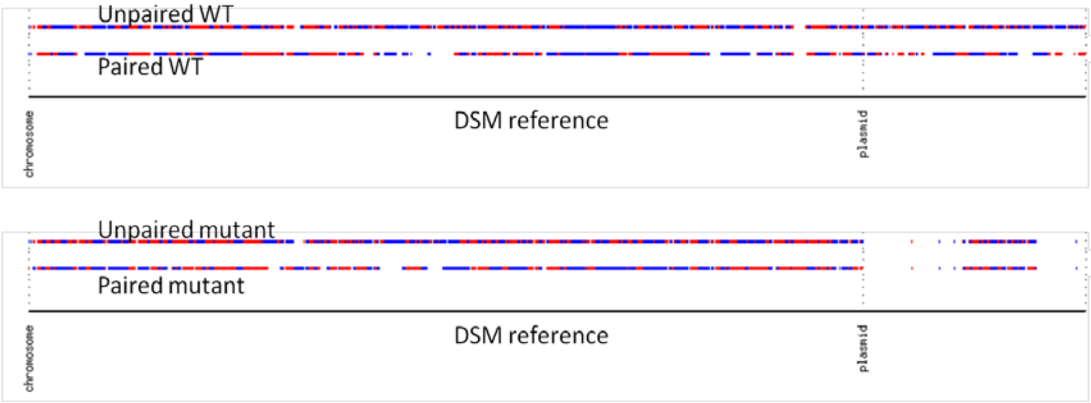


Figure 6.17. Comparison of contigs resulting from unpaired and paired assemblies for both strains. Note that the contigs from the paired assembly are fewer and longer, but the alignment is more fragmented.

6.6 Future directions

With the availability of at least three sequences for the *Streptomyces clavuligerus* WT strain, it should be possible to close the gaps still present in the so far most complete sequence, that from DSM.

Some differences were detected for the WT strain when compared to the DSM sequence. Although next generation sequencing technologies have a higher error rate, they also provide extremely deep coverage, and as shown in this work, the identification of mutations supported by 100% of the reads covering that position is possible. A criteria as stringent as the one used in this work could be applied and differences in sequence supported by 100% of the reads could be further investigated.

The impact of the mutation upstream of *ccaR* is currently being investigated. So far it seems to point to CcaR having a high impact on the intermediate steps in the clavulanic acid production pathway. As up to this moment the exact involvement of CcaR in clavulanic acid production has not been deciphered, this would represent a huge step forward in understanding the regulation of this potent β-lactamase inhibitor.

6.7 Concluding remarks

Streptomyces clavuligerus is a bacteria that produces several β-lactam antibiotics, including cephamycin C and clavulanic acid. Clavulanic acid is a potent β-lactamase inhibitor, used in combination with amoxicillin in the commercial product Augmentin^{TM}. In this project we analyzed a *Streptomyces clavuligerus* high producer strain and compared it to the WT strain. The mutant strain produces 2.5-3 times more clavulanic acid than the WT strain. The genome-wide mutation screening revealed that the mutant strain presents a smaller plasmid, resulting from the deletion of a region of 804 Kb at the 5' end of the plasmid, and of 387 Kb at the 3' end of the plasmid. Other important differences included a non silent mutation within the *cmcH* gene, which could block a step in the cephamycin C production pathway. The mutation most likely responsible for the high producing clavulanic acid phenotype in the mutant strain is a single nucleotide change located in the upstream region of the positive regulator *ccaR*. CcaR has an impact not only in clavulanic acid production, but also on cephamycin C production.

Chapter 7
Summary and Concluding Remarks

The study of microorganisms has made great strides over the years. By using different tools and approaches scientists and engineers have contributed to the understanding of how these complex biological systems work. In this work the focus was in particular in bacteria that produce secondary metabolites. More than 20,000 microbial seconday metabolites are known, and in nature they can act as antibiotics, toxins, ionophores, bioregulators and as signaling molecules (Marinelli 2009). In addition to their use as antibiotics in human and animal medicine, secondary metabolites have found use as antitumor agents, immunosupressants, hypocholesterolemic agents, antimigraine agents, enzyme inhibitors, and antiparasitic agents (Demain 1999). Secondary metabolites are usually produced by microorganisms in their late growth phase, and in culture secondary metabolite production is greatly influenced by the nutrients in the culture media (Ruiz, Chavez et al. 2010).

Most secondary metabolites are produced by filamentous fungi and by actinomycetes. In this work, regulation of the model organism for actinomycetes, *Streptomyces coelicolor*, was studied using systems approaches. As research has been conducted in this organism for more than 60 years, a multitude of information regarding its famous colored antibiotics has been accumulated. The availability of the genome sequence for this organism in 2000, lead to the construction of microarrays for probing its transcriptome by several research groups, including ours. A stage was then reached in which the number of accumulated transcriptome data could be analyzed as a whole, instead of as individual sets. Thus we used transcriptome data at the operon level to reconstruct the whole-genome regulatory network of *Streptomyces coelicolor*. The networks were assessed for the presence of known interactions in them, as well as for their enrichment in functional classes. The network prediction was combined with the identification of consensus sequences in the upstream region of the network members (Chapter 4). In twenty network modules, two of the features used for assessment, functional enrichment and presence of a consensus sequence in all of its members, were satisfied. These network modules are the most likely to contain true interactions, and thus provide a wealth of information as starting point for future experimental studies.

The next chapter of this thesis (Chapter 5) dealt with the curdlan producer *Agrobacterium* sp. ATCC 31749. Curdlan mimics secondary metabolite production. Curdlan production in *Agrobacterium* sp. ATCC 31749 is triggered by nitrogen starvation (Laroche and Michaud 2007). Curdlan is used in the food industry as an additive to modify the physical properties of products and as ingredient to develop new products (Miwa, Nakao et al. 1993). The study of *Agrobacterium* sp. ATCC 31749 was at a completely different level compared to that of *Streptomyces coelicolor*. In the case of *Agrobacterium* sp. ATCC 31749 no genomic resources existed and thus have to be created. A draft sequence for this organism was assembled from next generation sequencing. The draft sequence was compared to that of other close organisms, in particular to *Agrobacterium tumefaciens* C58, to which it was found to be extremely similar at the genomic level. Open reading frames were predicted and annotated and used as the basis for the construction by our collaborators at the Georgia Institute of Technology of an oligonucleotide microarray. The microarray is now being used to study *Agrobacterium* sp. ATCC 31749 and mutant strains created from the information generated by this work.

Next generation sequencing was also used to compare the genomes of two strains of *Streptomyces clavuligerus* (Chapter 6). *Streptomyces clavuligerus* produces two β-lactam antibiotics: cephamycin C and clavulanic acid. In this work we identified the differences between a wild type strain and a clavulanic acid high producer strain. The comparison was facilitated by the recent availability of the genome of the wild type strain *Streptomyces clavuligerus* ATCC 27064. The comparison revealed a heterogeneous population in the case of the mutant strain. Most of the cells in the population sequenced have lost two regions of 804 and 387 Kb in the extremities of the giant plasmid. Members of the *Streptomyces* genus are well known for the instability of their linear genomes, especially at the extremities (Leblond and Decaris 1994). A non-silent mutation was detected in *cmcH*, a key enzyme in the cephamycin pathway. A single point mutation was also detected in the promoter region of the regulator *ccaR*. CcaR has an effect on the production of both antibiotics, cephamycin C and clavulanic acid. Although binding of CcaR to genes in the cephamycin C cluster has been proven, the role of CcaR in the regulation of clavulanic acid is still unclear.

In this thesis, large volumes of data were used to study the regulation of secondary metabolites in different bacteria at different levels. The availability of the genome sequence for an organism pushes research to new levels of inquiry, and is the triggering point for systems approaches. Next generation sequencing technologies are increasing the number of organisms that can now be studied at an "omic" level. Vast volumes of data are now generated by transcriptomic studies of strains under diverse conditions and even that data can be integrated and analyzed as a whole. The resulting information can be integrated with biological understanding previously obtained by studying one gene at a time. Closing the cycle, the systems approaches provide targets for further study using traditional approaches. This work exemplified how we are integrating biological information from diverse levels in our quest to decipher amazingly complex microorganisms.

References

Aceti, D. J. and W. C. Champness (1998). "Transcriptional regulation of Streptomyces coelicolor pathway-specific antibiotic regulators by the absA and absB loci." J Bacteriol **180**(12): 3100-6.

Adachi, N. and M. R. Lieber (2002). "Bidirectional gene organization: a common architectural feature of the human genome." Cell **109**(7): 807-9.

Aharonowitz, Y. and A. L. Demain (1979). "Nitrogen nutrition and regulation of cephalosporin production in Streptomyces clavuligerus." Can J Microbiol **25**(1): 61-7.

Allardet-Servent, A., S. Michaux-Charachon, et al. (1993). "Presence of one linear and one circular chromosome in the Agrobacterium tumefaciens C58 genome." J Bacteriol **175**(24): 7869-74.

Alm, R. A., L. S. Ling, et al. (1999). "Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen Helicobacter pylori." Nature **397**(6715): 176-80.

Ansorge, W. J. (2009). "Next-generation DNA sequencing techniques." N Biotechnol **25**(4): 195-203.

Apel, A. K., A. Sola-Landa, et al. (2007). "Phosphate control of phoA, phoC and phoD gene expression in Streptomyces coelicolor reveals significant differences in binding of PhoP to their promoter regions." Microbiology **153**(Pt 10): 3527-37.

Arias, P., M. A. Fernandez-Moreno, et al. (1999). "Characterization of the pathway-specific positive transcriptional regulator for actinorhodin biosynthesis in Streptomyces coelicolor A3(2) as a DNA-binding protein." J Bacteriol **181**(22): 6958-68.

Bailey, T. L. and M. Gribskov (1998). "Combining evidence using p-values: application to sequence homology searches." Bioinformatics **14**(1): 48-54.

Bailey, T. L., N. Williams, et al. (2006). "MEME: discovering and analyzing DNA and protein sequence motifs." Nucleic Acids Res **34**(Web Server issue): W369-73.

Basso, K., A. A. Margolin, et al. (2005). "Reverse engineering of regulatory networks in human B cells." Nat Genet **37**(4): 382-90.

Beck, C. F. and R. A. Warren (1988). "Divergent promoters, a common form of gene organization." Microbiol Rev **52**(3): 318-26.

Bentley, D. R., S. Balasubramanian, et al. (2008). "Accurate whole human genome sequencing using reversible terminator chemistry." Nature **456**(7218): 53-9.

Bentley, S. D., K. F. Chater, et al. (2002). "Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2)." Nature **417**(6885): 141-147.

Berriman, M. and K. Rutherford (2003). "Viewing and annotating sequence data with Artemis." Brief Bioinform **4**(2): 124-32.

Besemer, J. and M. Borodovsky (2005). "GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses." Nucleic Acids Res **33**(Web Server issue): W451-4.

Besemer, J., A. Lomsadze, et al. (2001). "GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions." Nucleic Acids Res **29**(12): 2607-18.

Bignell, D. R., K. Tahlan, et al. (2005). "Expression of ccaR, encoding the positive activator of cephamycin C and clavulanic acid production in Streptomyces clavuligerus, is dependent on bldG." Antimicrob Agents Chemother **49**(4): 1529-41.

Binnewies, T. T., Y. Motro, et al. (2006). "Ten years of bacterial genome sequencing: comparative-genomics-based discoveries." Funct Integr Genomics **6**(3): 165-85.

Birch, J. R. and A. J. Racher (2006). "Antibody production." Adv Drug Deliv Rev **58**(5-6): 671-85.

Bird, A. E., J. M. Bellis, et al. (1982). "Spectrophotometric Assay of Clavulanic Acid by Reaction with Imidazole." Analyst **107**: 1241-1245.

Borodovsky, M. and J. McIninch (1993). "GENMARK: Parallel gene recognition for both DNA strands." Computers & Chemistry **17**(2): 123-133.

Brakhage, A. A., Q. Al-Abdallah, et al. (2005). "Evolution of [beta]-lactam biosynthesis genes and recruitment of trans-acting factors." Phytochemistry **66**(11): 1200-1210.

Brogden, R. N., A. Carmine, et al. (1981). "Amoxycillin/clavulanic acid: a review of its antibacterial activity, pharmacokinetics and therapeutic use." Drugs **22**(5): 337-62.

Brown, A. G., D. Butterworth, et al. (1976). "Naturally-occurring beta-lactamase inhibitors with antibacterial activity." J Antibiot (Tokyo) **29**(6): 668-9.

Bucca, G., E. Laing, et al. (2009). "Development and application of versatile high density microarrays for genome-wide analysis of Streptomyces coelicolor: characterization of the HspR regulon." Genome Biol **10**(1): R5.

Burrows, M. and D. J. Wheeler (1994). A block-sorting lossless data compression algorithm. Palo Alto, Digital Equipment Corporation.

Butler, J., I. MacCallum, et al. (2008). "ALLPATHS: de novo assembly of whole-genome shotgun microreads." Genome Res **18**(5): 810-20.

Butte, A. J. and I. S. Kohane (2000). "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements." Pac Symp Biocomput: 418-29.

Butte, A. J., P. Tamayo, et al. (2000). "Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks." Proc Natl Acad Sci U S A **97**(22): 12182-6.

Buttner, M. J., K. F. Chater, et al. (1990). "Cloning, disruption, and transcriptional analysis of three RNA polymerase sigma factor genes of Streptomyces coelicolor A3(2)." J Bacteriol **172**(6): 3367-78.

Buttner, M. J. and C. G. Lewis (1992). "Construction and characterization of Streptomyces coelicolor A3(2) mutants that are multiply deficient in the nonessential hrd-encoded RNA polymerase sigma factors." J Bacteriol **174**(15): 5165-7.

Cane, D. E., C. T. Walsh, et al. (1998). "Harnessing the biosynthetic code: combinations, permutations, and mutations." Science **282**(5386): 63-8.

Carver, T., U. Bohme, et al. (2010). "BamView: viewing mapped read alignment data in the context of the reference sequence." Bioinformatics **26**(5): 676-7.

Chang, C. C. and C. J. Lin (2001). "Training nu-support vector classifiers: theory and algorithms." Neural Comput **13**(9): 2119-47.

Chang, H. M., M. Y. Chen, et al. (1996). "The cutRS signal transduction system of Streptomyces lividans represses the biosynthesis of the polyketide antibiotic actinorhodin." Mol Microbiol **21**(5): 1075-85.

Charaniya, S. (2008). Systems analysis of complex biological data for bioprocess enhancement. Department of Chemical Engineering and Materials Science, University of Minnesota, Department of Chemical Engineering and Materials Science. **Ph. D.**

Charaniya, S., H. Le, et al. (2010). "Mining manufacturing data for discovery of high productivity process characteristics." J Biotechnol **147**(3-4): 186-97.

Charaniya, S., S. Mehra, et al. (2007). "Transcriptome dynamics-based operon prediction and verification in Streptomyces coelicolor." Nucleic Acids Res **35**(21): 7222-36.

Chater, K. (2006). "Review. <i>Streptomyces</i> inside-out: a new perspective on the bacteria that provide us with antibiotics." Philosophical Transactions of the Royal Society B: Biological Sciences **361**(1469): 761-768.

Chater, K. F. (1993). "Genetics of differentiation in Streptomyces." Annu Rev Microbiol **47**: 685-713.

Chater, K. F. (2001). "Regulation of sporulation in Streptomyces coelicolor A3(2): a checkpoint multiplex?" Curr Opin Microbiol **4**(6): 667-73.

Chater, K. F. and G. Chandra (2006). "The evolution of development in Streptomyces analysed by genome comparisons." FEMS Microbiol Rev **30**(5): 651-72.

Chen, C. W., C. H. Huang, et al. (2002). "Once the circle has been broken: dynamics and evolution of Streptomyces chromosomes." Trends Genet **18**(10): 522-9.

Claessen, D., W. de Jong, et al. (2006). "Regulation of Streptomyces development: reach for the sky!" Trends Microbiol **14**(7): 313-9.

Claessen, D., I. Stokroos, et al. (2004). "The formation of the rodlet layer of streptomycetes is the result of the interplay between rodlins and chaplins." Mol Microbiol **53**(2): 433-43.

Coco, E. A., K. E. Narva, et al. (1991). "New classes of Streptomyces coelicolor A3(2) mutants blocked in undecylprodigiosin (Red) biosynthesis." Mol Gen Genet **227**(1): 28-32.

Cronn, R., A. Liston, et al. (2008). "Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology." Nucleic Acids Res **36**(19): e122.

De Bona, F., S. Ossowski, et al. (2008). "Optimal spliced alignments of short sequence reads." Bioinformatics **24**(16): i174-80.

Dekker, C. (2007). "Solid-state nanopores." Nat Nanotechnol **2**(4): 209-15.

Delcher, A. L., K. A. Bratke, et al. (2007). "Identifying bacterial genes and endosymbiont DNA with Glimmer." Bioinformatics **23**(6): 673-9.

Delcher, A. L., D. Harmon, et al. (1999). "Improved microbial gene identification with GLIMMER." Nucleic Acids Res **27**(23): 4636-41.

Delcher, A. L., S. Kasif, et al. (1999). "Alignment of whole genomes." Nucleic Acids Res **27**(11): 2369-76.

Delcher, A. L., A. Phillippy, et al. (2002). "Fast algorithms for large-scale genome alignment and comparison." Nucleic Acids Res **30**(11): 2478-83.

Demain, A. L. (1999). "Pharmaceutically active secondary metabolites of microorganisms." Appl Microbiol Biotechnol **52**(4): 455-63.

Demain, A. L. and A. Fang (2000). "The natural functions of secondary metabolites." Adv Biochem Eng Biotechnol **69**: 1-39.

Ding, H. and M. F. Hynes (2009). "Plasmid transfer systems in the rhizobia." Can J Microbiol **55**(8): 917-27.

Do, J. H. and D. K. Choi (2006). "Computational approaches to gene prediction." J Microbiol **44**(2): 137-44.

Dohm, J. C., C. Lottaz, et al. (2007). "SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing." Genome Res **17**(11): 1697-706.

Donadio, Sosio, et al. (2002). "Impact of the first Streptomyces genome sequence on the discovery and production of bioactive substances." Applied Microbiology and Biotechnology **60**(4): 377-380.

Drews, J. (2000). "Drug discovery: a historical perspective." Science **287**(5460): 1960-4.

Duggleby, H. J., S. P. Tolley, et al. (1995). "Penicillin acylase has a single-amino-acid catalytic centre." Nature **373**(6511): 264-8.

Eaves, H. L. and Y. Gao (2009). "MOM: maximum oligonucleotide mapping." Bioinformatics **25**(7): 969-70.

Egan, E. S., M. A. Fogel, et al. (2005). "Divided genomes: negotiating the cell cycle in prokaryotes with multiple chromosomes." Mol Microbiol **56**(5): 1129-38.

Eid, J., A. Fehr, et al. (2009). "Real-time DNA sequencing from single polymerase molecules." Science **323**(5910): 133-8.

Faith, J. J., B. Hayete, et al. (2007). "Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles." PLoS Biol **5**(1): e8.

Farrand, S. K., P. B. Van Berkum, et al. (2003). "Agrobacterium is a definable genus of the family Rhizobiaceae." Int J Syst Evol Microbiol **53**(Pt 5): 1681-7.

Fernandez-Moreno, M. A., J. L. Caballero, et al. (1991). "The act cluster contains regulatory and antibiotic export genes, direct targets for translational control by the bldA tRNA gene of Streptomyces." Cell **66**(4): 769-80.

Fleischmann, R. D., M. D. Adams, et al. (1995). "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd." Science **269**(5223): 496-512.

Foulstone, M. and C. Reading (1982). "Assay of amoxicillin and clavulanic acid, the components of Augmentin, in biological fluids with high-performance liquid chromatography." Antimicrob Agents Chemother **22**(5): 753-62.

Gage, D. J. (2004). "Infection and invasion of roots by symbiotic, nitrogen-fixing rhizobia during nodulation of temperate legumes." Microbiol Mol Biol Rev **68**(2): 280-300.

Geddes, A. M., K. P. Klugman, et al. (2007). "Introduction: historical perspective and development of amoxicillin/clavulanate." International Journal of Antimicrobial Agents **30**(Supplement 2): 109-112.

Gehring, A. M., N. J. Yoo, et al. (2001). "RNA polymerase sigma factor that blocks morphological differentiation by Streptomyces coelicolor." J Bacteriol **183**(20): 5991-6.

Ghai, S. K. (1981). "Compositional studies on succinoglycan-like extracellular water-soluble Rhizobium polysaccharides." Acta Microbiol Pol **30**(2): 133-41.

Goodner, B., G. Hinkle, et al. (2001). "Genome sequence of the plant pathogen and biotechnology agent Agrobacterium tumefaciens C58." Science **294**(5550): 2323-8.

Gordon, N. D., G. L. Ottaviano, et al. (2008). "Secreted-protein response to sigmaU activity in Streptomyces coelicolor." J Bacteriol **190**(3): 894-904.

Gottelt, M., S. Kol, et al. "Deletion of a regulatory gene within the cpk gene cluster reveals novel antibacterial activity in Streptomyces coelicolor A3(2)." Microbiology **156**(Pt 8): 2343-53.

Grandvalet, C., V. de Crecy-Lagard, et al. (1999). "The ClpB ATPase of Streptomyces albus G belongs to the HspR heat shock regulon." Mol Microbiol **31**(2): 521-32.

Grynberg, M. and A. Godzik (2004). "NERD: a DNA processing-related domain present in the anthrax virulence plasmid, pXO1." Trends Biochem Sci **29**(3): 106-10.

Guo, F. B., H. Y. Ou, et al. (2003). "ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes." Nucleic Acids Res **31**(6): 1780-9.

Hall, N. (2007). "Advanced sequencing technologies and their wider impact in microbiology." J Exp Biol **210**(Pt 9): 1518-25.

Helmann, J. D. (2002). "The extracytoplasmic function (ECF) sigma factors." Adv Microb Physiol **46**: 47-110.

Hernandez, D., P. Francois, et al. (2008). "De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer." Genome Res **18**(5): 802-9.

Hesketh, A., H. Kock, et al. (2009). "The role of absC, a novel regulatory gene for secondary metabolism, in zinc-dependent antibiotic production in Streptomyces coelicolor A3(2)." Mol Microbiol **74**(6): 1427-44.

Higgens, C. E. and R. E. Kastner (1971). "Streptomyces clavuligerus sp. no., a beta-Lactam Antibiotic Producer." International Journal of Systematic Bateriology **21**(4): 326-331.

Hocquette, J. F. (2005). "Where are we in genomics?" J Physiol Pharmacol **56 Suppl 3**: 37-70.

Hojati, Z., C. Milne, et al. (2002). "Structure, biosynthetic origin, and engineered biosynthesis of calcium-dependent antibiotics from Streptomyces coelicolor." Chem Biol **9**(11): 1175-87.

Hojati, Z., C. Milne, et al. (2002). "Structure, Biosynthetic Origin, and Engineered Biosynthesis of Calcium-Dependent Antibiotics from Streptomyces coelicolor." Chemistry & Biology **9**(11): 1175-1187.

Holtorf, H., M. C. Guitton, et al. (2002). "Plant functional genomics." Naturwissenschaften **89**(6): 235-49.

Homer, N., B. Merriman, et al. (2009). "Local alignment of two-base encoded DNA sequence." BMC Bioinformatics **10**: 175.

Hong, H.-J., M. I. Hutchings, et al. (2004). "Characterization of an inducible vancomycin resistance system in Streptomyces coelicolor reveals a novel gene (vanK) required for drug resistance." Molecular Microbiology **52**(4): 1107-1121.

Hopwood, D. A. (2006). "Soil To Genomics: The Streptomyces Chromosome." Annual Review of Genetics **40**(1): 1-23.

Hopwood, D. A., K. F. Chater, et al. (1973). "Advances in Streptomyces coelicolor genetics." Bacteriol Rev **37**(3): 371-405.

Hossain, M. S., N. Azimi, et al. (2009). "Crystallizing short-read assemblies around seeds." BMC Bioinformatics **10 Suppl 1**: S16.

Hutchings, M. I., P. A. Hoskisson, et al. (2004). "Sensing and responding to diverse extracellular signals? Analysis of the sensor kinases and response regulators of Streptomyces coelicolor A3(2)." Microbiology **150**(Pt 9): 2795-806.

Hyatt, D., G. L. Chen, et al. (2010). "Prodigal: prokaryotic gene recognition and translation initiation site identification." BMC Bioinformatics **11**: 119.

Ishizuka, H., S. Horinouchi, et al. (1992). "A putative two-component regulatory system involved in secondary metabolism in Streptomyces spp." J Bacteriol **174**(23): 7585-94.

Jagodzinski, P. P., R. Wiaderkiewicz, et al. (1994). "Mechanism of the inhibitory effect of curdlan sulfate on HIV-1 infection in vitro." Virology **202**(2): 735-45.

Jain, E. and A. Kumar (2008). "Upstream processes in antibody production: evaluation of critical parameters." Biotechnol Adv **26**(1): 46-72.

Jeck, W. R., J. A. Reinhardt, et al. (2007). "Extending assembly of short DNA sequences to handle error." Bioinformatics **23**(21): 2942-4.

Jensen, S. E. and A. S. Paradkar (1999). "Biosynthesis and molecular genetics of clavulanic acid." Antonie Van Leeuwenhoek **75**(1-2): 125-33.

Jensen, S. E., A. S. Paradkar, et al. (2004). "Five additional genes are involved in clavulanic acid biosynthesis in Streptomyces clavuligerus." Antimicrob Agents Chemother **48**(1): 192-202.

Jiang, H. and W. H. Wong (2008). "SeqMap: mapping massive amount of oligonucleotides to the genome." Bioinformatics **24**(20): 2395-6.

Jumas-Bilak, E., S. Michaux-Charachon, et al. (1998). "Unconventional genomic organization in the alpha subgroup of the Proteobacteria." J Bacteriol **180**(10): 2749-55.

Kallifidas, D., B. Pascoe, et al. (2010). "The zinc-responsive regulator Zur controls expression of the coelibactin gene cluster in Streptomyces coelicolor." J Bacteriol **192**(2): 608-11.

Kantardjieff, A. (2009). Transcriptome analysis in mammalian cell culture: applications in process development and characterization. Department of Chemical Engineering and Materials Science, University of Minnesota, Department of Chemical Engineering and Materials Science. **Ph. D.**

Karnezis, T., V. C. Epa, et al. (2003). "Topological characterization of an inner membrane (1-->3)-beta-D-glucan (curdlan) synthase from Agrobacterium sp. strain ATCC31749." Glycobiology **13**(10): 693-706.

Karnezis, T., H. C. Fisher, et al. (2002). "Cloning and characterization of the phosphatidylserine synthase gene of Agrobacterium sp. strain ATCC 31749 and effect of its inactivation on production of high-molecular-mass (1-->3)-beta-D-glucan (curdlan)." J Bacteriol **184**(15): 4114-23.

Kasianowicz, J. J., E. Brandin, et al. (1996). "Characterization of individual polynucleotide molecules using a membrane channel." Proc Natl Acad Sci U S A **93**(24): 13770-3.

Kieser, T., M. J. Bibb, et al. (2000). Practical Streptomyces Genetics. Norwich, England, John Innes Foundation.

Kim, H. S., Y. J. Lee, et al. (2004). "Cloning and characterization of a gene encoding the gamma-butyrolactone autoregulator receptor from Streptomyces clavuligerus." Arch Microbiol **182**(1): 44-50.

Kim, H. S. and Y. I. Park (2008). "Isolation and identification of a novel microorganism producing the immunosuppressant tacrolimus." J Biosci Bioeng **105**(4): 418-21.

Koehn, F. E. and G. T. Carter (2005). "The evolving role of natural products in drug discovery." Nat Rev Drug Discov **4**(3): 206-20.

Komatsu, M., H. Takano, et al. (2006). "Proteins encoded by the conservon of Streptomyces coelicolor A3(2) comprise a membrane-associated heterocomplex that resembles eukaryotic G protein-coupled regulatory system." Mol Microbiol **62**(6): 1534-46.

Korbel, J. O., L. J. Jensen, et al. (2004). "Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs." Nat Biotechnol **22**(7): 911-7.

Kormanec, J. and M. Farkasovsky (1993). "Differential expression of principal sigma factor homologues of Streptomyces aureofaciens correlates with the developmental stage." Nucleic Acids Res **21**(16): 3647-52.

Kormanec, J., M. Farkasovsky, et al. (1992). "Four genes in Streptomyces aureofaciens containing a domain characteristic of principal sigma factors." Gene **122**(1): 63-70.

Kurtz, S., A. Phillippy, et al. (2004). "Versatile and open software for comparing large genomes." Genome Biol **5**(2): R12.

Kyrpides, N. C. (2009). "Fifteen years of microbial genomics: meeting the challenges and fulfilling the dream." Nat Biotechnol **27**(7): 627-32.

Kyung, Y. S., W. S. Hu, et al. (2001). "Analysis of temporal and spatial expression of the CcaR regulatory element in the cephamycin C biosynthetic pathway using green fluorescent protein." Mol Microbiol **40**(3): 530-41.

Lakey, J. H., E. J. Lea, et al. (1983). "A new channel-forming antibiotic from Streptomyces coelicolor A3(2) which requires calcium for its activity." J Gen Microbiol **129**(12): 3565-73.

Langmead, B., C. Trapnell, et al. (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." Genome Biol **10**(3): R25.

Laroche, C. and P. Michaud (2007). "New developments and prospective applications for beta (1,3) glucans." Recent Pat Biotechnol **1**(1): 59-73.

Leblond, P. and B. Decaris (1994). "New insights into the genetic instability of streptomyces." FEMS Microbiol Lett **123**(3): 225-32.

Li, H. and R. Durbin (2009). "Fast and accurate short read alignment with Burrows-Wheeler transform." Bioinformatics **25**(14): 1754-60.

Li, H., B. Handsaker, et al. (2009). "The Sequence Alignment/Map format and SAMtools." Bioinformatics **25**(16): 2078-9.

Li, H. and N. Homer (2010). "A survey of sequence alignment algorithms for next-generation sequencing." Brief Bioinform.

Li, H., J. Ruan, et al. (2008). "Mapping short DNA sequencing reads and calling variants using mapping quality scores." Genome Res **18**(11): 1851-8.

Li, J. W. and J. C. Vederas (2009). "Drug discovery and natural products: end of an era or an endless frontier?" Science **325**(5937): 161-5.

Li, R., W. Fan, et al. (2010). "The sequence and de novo assembly of the giant panda genome." Nature **463**(7279): 311-7.

Li, R., Y. Li, et al. (2008). "SOAP: short oligonucleotide alignment program." Bioinformatics **24**(5): 713-4.

Li, R., C. Yu, et al. (2009). "SOAP2: an improved ultrafast tool for short read alignment." Bioinformatics **25**(15): 1966-7.

Li, R., H. Zhu, et al. (2010). "De novo assembly of human genomes with massively parallel short read sequencing." Genome Res **20**(2): 265-72.

Li, W., D. Raoult, et al. (2009). "Bacterial strain typing in the genomic era." FEMS Microbiol Rev **33**(5): 892-916.

Lian, W., K. P. Jayapal, et al. (2008). "Genome-wide transcriptome analysis reveals that a pleiotropic antibiotic regulator, AfsS, modulates nutritional stress response in Streptomyces coelicolor A3(2)." BMC Genomics **9**: 56.

Liras, P. (1999). "Biosynthesis and molecular genetics of cephamycins. Cephamycins produced by actinomycetes." Antonie Van Leeuwenhoek **75**(1-2): 109-24.

Liras, P. and A. L. Demain (2009). "Chapter 16. Enzymology of beta-lactam compounds with cephem structure produced by actinomycete." Methods Enzymol **458**: 401-29.

Liras, P., J. P. Gomez-Escribano, et al. (2008). "Regulatory mechanisms controlling antibiotic production in Streptomyces clavuligerus." J Ind Microbiol Biotechnol **35**(7): 667-76.

Liras, P. and J. F. Martin (2005). "Assay Methods for Detection and Quantification of Antimicrobial Metabolites Produced by Streptomyces clavuligerus." Microbial Processes and Products: 149-153.

Liras, P. and A. RodrÃ-guez-Garcia (2000). "Clavulanic acid, a Î²-lactamase inhibitor: biosynthesis and molecular genetics." Applied Microbiology & Biotechnology **54**(4): 467-475.

Lukashin, A. V. and M. Borodovsky (1998). "GeneMark.hmm: new solutions for gene finding." Nucleic Acids Res **26**(4): 1107-15.

Maccallum, I., D. Przybylski, et al. (2009). "ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads." Genome Biol **10**(10): R103.

MacLean, D., J. D. Jones, et al. (2009). "Application of 'next-generation' sequencing technologies to microbial genetics." Nat Rev Microbiol **7**(4): 287-96.

Makitrynskyy, R., Y. Rebets, et al. (2010). "Genetic factors that influence moenomycin production in streptomycetes." J Ind Microbiol Biotechnol **37**(6): 559-66.

Marcos, A. T., S. Gutierrez, et al. (1995). "Three genes hrdB, hrdD and hrdT of Streptomyces griseus IMRU 3570, encoding sigma factor-like proteins, are differentially expressed under specific nutritional conditions." Gene **153**(1): 41-8.

Mardis, E. R. (2008). "Next-generation DNA sequencing methods." Annu Rev Genomics Hum Genet **9**: 387-402.

Margolin, A. A., I. Nemenman, et al. (2006). "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context." BMC Bioinformatics **7 Suppl 1**: S7.

Margolin, A. A., K. Wang, et al. (2006). "Reverse engineering cellular networks." Nat Protoc **1**(2): 662-71.

Margulies, M., M. Egholm, et al. (2005). "Genome sequencing in microfabricated high-density picolitre reactors." Nature **437**(7057): 376-80.

Marinelli, F. (2009). "Chapter 2. From microbial products to novel drugs that target a multitude of disease indications." Methods Enzymol **458**: 29-58.

Markowitz, V. M., N. N. Ivanova, et al. (2008). Using IMG. Comparative Analysis with the Integrated Microbial Genomes System, Genome Biology Program. Department of Energy Joint Genome Institute.

Biological data Management and Technology Center. Lawrence berkeley National Laboratory.**:** 1-55.

Markowitz, V. M., E. Szeto, et al. (2008). "The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions." Nucleic Acids Res **36**(Database issue): D528-33.

Martin, J. F., J. Casqueiro, et al. (2005). "Secretion systems for secondary metabolites: how producer cells send out messages of intercellular communication." Curr Opin Microbiol **8**(3): 282-93.

McCutcheon, J. P., B. R. McDonald, et al. (2009). "Origin of an alternative genetic code in the extremely small and GC-rich genome of a bacterial symbiont." PLoS Genet **5**(7): e1000565.

McIntosh, M., B. A. Stone, et al. (2005). "Curdlan and other bacterial (1-->3)-beta-D-glucans." Appl Microbiol Biotechnol **68**(2): 163-73.

McKenzie, N. L. and J. R. Nodwell (2007). "Phosphorylated AbsA2 negatively regulates antibiotic production in Streptomyces coelicolor through interactions with pathway-specific regulatory gene promoters." J Bacteriol **189**(14): 5284-92.

McPherson, J. D. (2009). "Next-generation gap." Nat Methods **6**(11 Suppl): S2-5.

Medema, M. H., A. Trefzer, et al. (2010). "The sequence of a 1.8-mb bacterial linear plasmid reveals a rich evolutionary reservoir of secondary metabolic pathways." Genome Biol Evol **2**: 212-24.

Medini, D., D. Serruto, et al. (2008). "Microbiology in the post-genomic era." Nat Rev Microbiol **6**(6): 419-30.

Meyer, P. E., K. Kontos, et al. (2007). "Information-theoretic inference of large transcriptional regulatory networks." EURASIP J Bioinform Syst Biol: 79879.

Meyer, P. E., F. Lafitte, et al. (2008). "minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information." BMC Bioinformatics **9**: 461.

Miguelez, E. M., C. Hardisson, et al. (2000). "Streptomycetes: a new model to study cell death." Int Microbiol **3**(3): 153-8.

Miller, J. R., S. Koren, et al. (2010). "Assembly algorithms for next-generation sequencing data." Genomics **95**(6): 315-27.

Miwa, M., Y. Nakao, et al. (1993). Food applications of curdlan. Food hydrocolloids: structures, properties, and functions. K. Nishinari and E. Doi. New York, Plenum Press**:** 119-125.

Miyadoh, S. (1993). "Research on antibiotic screening in Japan over the last decade: a producing microorganisms approach." Actinomycetologica **9**: 100-106.

Moore, B. S. and J. Piel (2000). "Engineering biodiversity with type II polyketide synthase genes." Antonie Van Leeuwenhoek **78**(3-4): 391-8.

Nakabachi, A., A. Yamashita, et al. (2006). "The 160-kilobase genome of the bacterial endosymbiont Carsonella." Science **314**(5797): 267.

Nakanishi, I., K. Kimura, et al. (1976). "Demonstration of curdlan-type polysaccharide and some other beta-1,3-glucan in microorganisms with aniline blue." J Gen Appl Microbiol **22**(1): 1-11.

Nishi, T., T. Ikemura, et al. (2005). "GeneLook: a novel ab initio gene identification system suitable for automated annotation of prokaryotic sequences." Gene **346**: 115-25.

Nusbaum, C., T. K. Ohsumi, et al. (2009). "Sensitive, specific polymorphism discovery in bacteria using massively parallel sequencing." Nat Methods **6**(1): 67-9.

O'Connor, T. J., P. Kanellis, et al. (2002). "The ramC gene is required for morphogenesis in Streptomyces coelicolor and expressed in a cell type-specific manner under the direct control of RamR." Mol Microbiol **45**(1): 45-57.

Ochi, K. (2007). "From microbial differentiation to ribosome engineering." Biosci Biotechnol Biochem **71**(6): 1373-86.

Okami, Y. and K. Hotta (1988). Search and discovery of new antibiotics. Actinomycetes in biotechnology. M. Goodfellow, S. T. Williams and M. Mordarski. San diego, Academic Press**:** 33-67.

Ostash, B., E. H. Doud, et al. (2009). "Complete characterization of the seventeen step moenomycin biosynthetic pathway." Biochemistry **48**(37): 8830-41.

Paget, M. S., V. Molle, et al. (2001). "Defining the disulphide stress response in Streptomyces coelicolor A3(2): identification of the sigmaR regulon." Mol Microbiol **42**(4): 1007-20.

Paradkar, A., A. Trefzer, et al. (2003). "Streptomyces genetics: a genomic perspective." Crit Rev Biotechnol **23**(1): 1-27.

Paradkar, A. S., K. A. Aidoo, et al. (1998). "A pathway-specific transcriptional activator regulates late steps of clavulanic acid biosynthesis in Streptomyces clavuligerus." Mol Microbiol **27**(4): 831-43.

Paradkar, A. S. and S. E. Jensen (1995). "Functional analysis of the gene encoding the clavaminate synthase 2 isoenzyme involved in clavulanic acid biosynthesis in Streptomyces clavuligerus." J Bacteriol **177**(5): 1307-14.

Perez-Llarena, F. J., P. Liras, et al. (1997). "A regulatory gene (ccaR) required for cephamycin and clavulanic acid production in Streptomyces clavuligerus: amplification results in overproduction of both beta-lactam compounds." J Bacteriol **179**(6): 2053-9.

Petrickova, K. and M. Petricek (2003). "Eukaryotic-type protein kinases in Streptomyces coelicolor: variations on a common theme." Microbiology **149**(7): 1609-1621.

Pevzner, P. A., H. Tang, et al. (2001). "An Eulerian path approach to DNA fragment assembly." Proc Natl Acad Sci U S A **98**(17): 9748-53.

Phillips, K. R. and H. G. Lawford (1983). "Curdlan: its properties and production in batch and continuous fermentations." Progress in industial microbiology **18**: 201-229.

Phillips, K. R., J. Pik, et al. (1983). "Production of curdlan-type polysaccharide by Alcaligenes faecalis in batch and continuous culture." Can J Microbiol **29**(10): 1331-8.

Pitzschke, A. and H. Hirt (2010). "New insights into an old story: Agrobacterium-induced tumour formation in plants by plant transformation." Embo J **29**(6): 1021-32.

Porcar, M. (2010). "Beyond directed evolution: Darwinian selection as a tool for synthetic biology." Syst Synth Biol **4**(1): 1-6.

Prell, J. and P. Poole (2006). "Metabolic changes of rhizobia in legume nodules." Trends Microbiol **14**(4): 161-8.

Prufer, K., U. Stenzel, et al. (2008). "PatMaN: rapid alignment of short sequences to large databases." Bioinformatics **24**(13): 1530-1.

Raju, T. N. (1999). "The Nobel chronicles. 1952: Selman Abraham Waksman (1888-1973)." Lancet **353**(9163): 1536.

Rathore, A. S. (2009). "Roadmap for implementation of quality by design (QbD) for biotechnology products." Trends Biotechnol **27**(9): 546-53.

Read, E. K., J. T. Park, et al. (2010). "Process analytical technology (PAT) for biopharmaceutical products: Part I. concepts and applications." Biotechnol Bioeng **105**(2): 276-84.

Rius, N. and A. L. Demain (1997). "Lysine epsilon-aminotransferase, the initial enzyme of cephalosporin biosynthesis in actinomycetes." J Microbiol Biotechnol **7**(2): 95-100.

Rodríguez-García, A., C. Barreiro, et al. (2007). "Genome-wide transcriptomic and proteomic analysis of the primary response to phosphate limitation in <B><I>Streptomyces coelicolor</I></B> M145 and in a Delta<B><I>phoP</I></B> mutant." PROTEOMICS **7**(14): 2410-2429.

Rodriguez-Garcia, A., M. Ludovice, et al. (1997). "Arginine boxes and the argR gene in Streptomyces clavuligerus: evidence for a clear regulation of the arginine pathway." Mol Microbiol **25**(2): 219-28.

Rothberg, J. M. and J. H. Leamon (2008). "The development and impact of 454 sequencing." Nat Biotechnol **26**(10): 1117-24.

Rothschild, M. F. and G. S. Plastow (2008). "Impact of genomics on animal agriculture and opportunities for animal health." Trends Biotechnol **26**(1): 21-5.

Rougemont, J., A. Amzallag, et al. (2008). "Probabilistic base calling of Solexa sequencing data." BMC Bioinformatics **9**: 431.

Ruffing, A. M. and R. R. Chen (2010). "Metabolic engineering of Agrobacterium sp. strain ATCC 31749 for production of an alpha-Gal epitope." Microb Cell Fact **9**: 1.

Ruiz, B., A. Chavez, et al. (2010). "Production of microbial secondary metabolites: regulation by the carbon source." Crit Rev Microbiol **36**(2): 146-67.

Rumble, S. M., P. Lacroute, et al. (2009). "SHRiMP: accurate mapping of short color-space reads." PLoS Comput Biol **5**(5): e1000386.

Rutherford, K., J. Parkhill, et al. (2000). "Artemis: sequence visualization and annotation." Bioinformatics **16**(10): 944-5.

Ryding, N. J., T. B. Anderson, et al. (2002). "Regulation of the Streptomyces coelicolor calcium-dependent antibiotic by absA, encoding a cluster-linked two-component system." J Bacteriol **184**(3): 794-805.

Sakakibara, Y. and B. C. Saha (2008). "Isolation of an operon involved in xylitol metabolism from a xylitol-utilizing Pantoea ananatis mutant." J Biosci Bioeng **106**(4): 337-44.

Salzberg, S. L., A. L. Delcher, et al. (1998). "Microbial gene identification using interpolated Markov models." Nucleic Acids Res **26**(2): 544-8.

Santamarta, I., M. T. Lopez-Garcia, et al. (2007). "Connecting primary and secondary metabolism: AreB, an IclR-like protein, binds the ARE(ccaR) sequence of S. clavuligerus and modulates leucine biosynthesis and cephamycin C and clavulanic acid production." Mol Microbiol **66**(2): 511-24.

Santamarta, I., R. Perez-Redondo, et al. (2005). "Different proteins bind to the butyrolactone receptor protein ARE sequence located upstream of the regulatory ccaR gene of Streptomyces clavuligerus." Mol Microbiol **56**(3): 824-35.

Santamarta, I., A. Rodriguez-Garcia, et al. (2002). "CcaR is an autoregulatory protein that binds to the ccaR and cefD-cmcI promoters of the cephamycin C-clavulanic acid cluster in Streptomyces clavuligerus." J Bacteriol **184**(11): 3106-13.

Saudagar, P. S. and R. S. Singhal (2004). "Fermentative production of curdlan." Appl Biochem Biotechnol **118**(1-3): 21-31.

Saudagar, P. S., S. A. Survase, et al. (2008). "Clavulanic acid: A review." Biotechnology Advances **26**(4): 335-351.

Schatz, M. C., C. Trapnell, et al. (2007). "High-throughput sequence alignment using Graphics Processing Units." BMC Bioinformatics **8**: 474.

Schena, M., D. Shalon, et al. (1995). "Quantitative monitoring of gene expression patterns with a complementary DNA microarray." Science **270**(5235): 467-70.

Schneiker, S., O. Perlova, et al. (2007). "Complete genome sequence of the myxobacterium Sorangium cellulosum." Nat Biotechnol **25**(11): 1281-9.

Servant, P. and P. Mazodier (2001). "Negative regulation of the heat shock response in Streptomyces." Arch Microbiol **176**(4): 237-42.

Shendure, J. and H. Ji (2008). "Next-generation DNA sequencing." Nat Biotechnol **26**(10): 1135-45.

Smith, A. D., Z. Xuan, et al. (2008). "Using quality scores and longer reads improves accuracy of Solexa read mapping." BMC Bioinformatics **9**: 128.

Smyth, G. K. (2004). "Linear models and empirical bayes methods for assessing differential expression in microarray experiments." Statistical applications in genetics and molecular biology **3**.

Sobczyk, A., A. Bellier, et al. (2002). "The lon gene, encoding an ATP-dependent protease, is a novel member of the HAIR/HspR stress-response regulon in actinomycetes." Microbiology **148**(Pt 6): 1931-7.

Sobral, B. W., R. J. Honeycutt, et al. (1991). "Electrophoretic separation of the three Rhizobium meliloti replicons." J Bacteriol **173**(16): 5173-80.

Sola-Landa, A., A. Rodriguez-Garcia, et al. (2005). "Binding of PhoP to promoters of phosphate-regulated genes in Streptomyces coelicolor: identification of PHO boxes." Mol Microbiol **56**(5): 1373-85.

Stasinopoulos, S. J., P. R. Fisher, et al. (1999). "Detection of two loci involved in (1-->3)-beta-glucan (curdlan) biosynthesis by Agrobacterium sp. ATCC31749, and comparative sequence analysis of the putative curdlan synthase gene." Glycobiology **9**(1): 31-41.

Studholme, D. J., S. D. Bentley, et al. (2004). "Bioinformatic identification of novel regulatory DNA sequence motifs in Streptomyces coelicolor." BMC Microbiol **4**: 14.

Sutherland, R. (1991). "Beta-lactamase inhibitors and reversal of antibiotic resistance." Trends Pharmacol Sci **12**(6): 227-32.

Takano, E., H. Kinoshita, et al. (2005). "A bacterial hormone (the SCB1) directly controls the expression of a pathway-specific regulatory gene in the cryptic type I polyketide biosynthetic gene cluster of Streptomyces coelicolor." Mol Microbiol **56**(2): 465-79.

Takano, E., T. Nihira, et al. (2000). "Purification and structural determination of SCB1, a gamma-butyrolactone that elicits antibiotic production in Streptomyces coelicolor A3(2)." J Biol Chem **275**(15): 11010-6.

Takano, E., M. Tao, et al. (2003). "A rare leucine codon in adpA is implicated in the morphological defect of bldA mutants of Streptomyces coelicolor." Molecular Microbiology **50**(2): 475-486.

Teixeira, A. P., R. Oliveira, et al. (2009). "Advances in on-line monitoring and control of mammalian cell cultures: Supporting the PAT initiative." Biotechnol Adv **27**(6): 726-32.

Tettelin, H., V. Masignani, et al. (2005). "Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome"." Proc Natl Acad Sci U S A **102**(39): 13950-5.

Thomas, S. H., R. D. Wagner, et al. (2008). "The mosaic genome of Anaeromyxobacter dehalogenans strain 2CP-C suggests an aerobic common ancestor to the delta-proteobacteria." PLoS One **3**(5): e2103.

Thompson, C. J., D. Fink, et al. (2002). "Principles of microbial alchemy: insights from the Streptomyces coelicolor genome sequence." Genome Biol **3**(7): REVIEWS1020.

Touzain, F., S. Schbath, et al. (2008). "SIGffRid: a tool to search for sigma factor binding sites in bacterial genomes using comparative approach and biologically driven statistics." BMC Bioinformatics **9**: 73.

Trepanier, N. K., S. E. Jensen, et al. (2002). "The positive activator of cephamycin C and clavulanic acid production in Streptomyces clavuligerus is mistranslated in a bldA mutant." Microbiology **148**(Pt 3): 643-56.

Tsao, S. W., B. A. Rudd, et al. (1985). "Identification of a red pigment from Streptomyces coelicolor A3(2) as a mixture of prodigiosin derivatives." J Antibiot (Tokyo) **38**(1): 128-31.

Turner, D. J., T. M. Keane, et al. (2009). "Next-generation sequencing of vertebrate experimental organisms." Mamm Genome.

Tusher, V. G., R. Tibshirani, et al. (2001). "Significance analysis of microarrays applied to the ionizing radiation response." Proc Natl Acad Sci U S A **98**(9): 5116-21.

van Keulen, G., D. A. Hopwood, et al. (2005). "Gas vesicles in actinomycetes: old buoys in novel habitats?" Trends Microbiol **13**(8): 350-4.

Wackett, L. P., J. A. Frias, et al. (2007). "Genomic and biochemical studies demonstrating the absence of an alkane-producing phenotype in Vibrio furnissii M1." Appl Environ Microbiol **73**(22): 7192-8.

Walsby, A. E. and P. G. Dunton (2006). "Gas vesicles in actinomycetes?" Trends Microbiol **14**(3): 99-100.

Wang, L., K. Tahlan, et al. (2004). "Transcriptional and translational analysis of the ccaR gene from Streptomyces clavuligerus." Microbiology **150**(Pt 12): 4137-45.

Wang, Q., L. Wan, et al. (2009). "Searching for bidirectional promoters in Arabidopsis thaliana." BMC Bioinformatics **10 Suppl 1**: S29.

Warren, R. L., G. G. Sutton, et al. (2007). "Assembling millions of short DNA sequences using SSAKE." Bioinformatics **23**(4): 500-1.

Welch, R. A., V. Burland, et al. (2002). "Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli." Proc Natl Acad Sci U S A **99**(26): 17020-4.

Werner, T. (2010). "Next generation sequencing in functional genomics." Brief Bioinform **11**(5): 499-511.

White, J. and M. Bibb (1997). "bldA dependence of undecylprodigiosin production in Streptomyces coelicolor A3(2) involves a pathway-specific regulatory cascade." J. Bacteriol. **179**(3): 627-633.

Wietzorrek, A. and M. Bibb (1997). "A novel family of proteins that regulates antibiotic production in streptomycetes appears to contain an OmpR-like DNA-binding fold." Mol Microbiol **25**(6): 1181-4.

Wilkinson, S. P. and A. Grove (2006). "Ligand-responsive transcriptional regulation by members of the MarR family of winged helix proteins." Curr Issues Mol Biol **8**(1): 51-62.

Willey, J. M., A. Willems, et al. (2006). "Morphogenetic surfactants and their role in the formation of aerial hyphae in Streptomyces coelicolor." Mol Microbiol **59**(3): 731-42.

Wood, D. W., J. C. Setubal, et al. (2001). "The genome of the natural genetic engineer Agrobacterium tumefaciens C58." Science **294**(5550): 2317-23.

Young, J. M., L. D. Kuykendall, et al. (2001). "A revision of Rhizobium Frank 1889, with an emended description of the genus, and the inclusion of all species of Agrobacterium Conn 1942 and Allorhizobium undicola de Lajudie et al. 1998 as new combinations: Rhizobium radiobacter, R. rhizogenes, R. rubi, R. undicola and R. vitis." Int J Syst Evol Microbiol **51**(Pt 1): 89-103.

Young, J. M., L. D. Kuykendall, et al. (2003). "Classification and nomenclature of Agrobacterium and Rhizobium." Int J Syst Evol Microbiol **53**(Pt 5): 1689-95.

Zerbino, D. R. and E. Birney (2008). "Velvet: algorithms for de novo short read assembly using de Bruijn graphs." Genome Res **18**(5): 821-9.

Zerbino, D. R., G. K. McEwen, et al. (2009). "Pebble and rock band: heuristic resolution of repeats and scaffolding in the velvet short-read de novo assembler." PLoS One **4**(12): e8407.

Appendix A Network modules containing a consensus sequence.

Table A.1. 414 network modules containing a consensus sequence

Network	*E*-value$_t$	*E*_value	Cistrons in network	Cistrons with motif	Consensus sequence
4	2.20E+01	9.00E-01	12	4	G[AC]C[GC][AT][CGT][GC]A[TAC][GC][ACT][CG][CAT][TA][CT]G[AGT][AT]G[CAT][ACGTA][GT][AGT][AT]G[GA][GAC][GA][AC][GT][CA][TA]C[GC][AGT][CA]G[AG]G[CT][TC]C[ACGTA]TC
6	1.30E+00	1.60E-01	10	3	[AT]TT[AT][GC][AG]T[TC][AG][TG]G[GA][CT][CG][CA][TC][AT][AT][TA]C[TA]AA[TA]
7	1.30E+00	1.20E+00	5	4	GTG[AC][ACGTA][AG][GC][AGT][ACT][CG][AG][GAC][TAG][CGT][AT][CT][ACG][AT][TAG][GT][AG][AG][CG][AC][TG][CG][AT][TA][GT][AT][TAG][CT][CAT][CAT][TG][GCT][CT][ACT][GAC]T
9	7.50E-02	4.10E-03	11	6	[AG][CA]TTC[TG]AA[AG][GA][TG]GT[CG]G[CG][CG][TA][TC]TGAT[GT][GC][CT]T
10	4.60E-01	2.10E-01	9	5	[CA][TG][AG][GC][AC][GAT][GAC][CA][TACGA][TC][GA][AC][ACT][TCG][AC][CG][GC][CAT][ACT][CAT][CT][TAG][GTC][AT][CG][CGA][TA][GA][CTA][GC][TG][TG][GTC][TA]T[GT][AG]
11	1.80E+01	1.80E-01	12	9	[TA]G[AT]TCA[CA][ACG][CAT]C[GT][GAT][TC]T
12	4.50E+01	9.40E+00	13	4	T[CGT][CG][CG][CG][GA][AC][AC][CG][GA]G[AGT][AG][CG][GAC][TG][CT]G[AC][GA]G[AT][TC][AGT][CA][CGT]GAA[CGT][CGT][CAT][AG][TG][CA][GT][GC][AGT]AC[AT]C
13	1.20E+01	1.90E+00	9	5	[TG][GTA][CG][GC][TA][CG]G[ACG][CG][GC][TA][CT]GACG[GA][AC][ACG]TC[GAT][ACT]C
14	1.60E+01	8.10E+00	12	9	TGA[CA][GT]TC[GA][TAC]
15	1.20E+01	5.60E+00	10	2	AAACGACGCA[AT]G[CT][AT][AG]C[AT]TGCGTCGTTT
16	7.60E+00	1.50E+00	13	5	T[GA][TA]ACG[CA][GC][TA][GAT][CTG][CTA][CT]A[TG]G[CA]
17	2.50E+01	2.50E-04	17	10	[AC][CG][GT][GC]CG[AG][AGC][GC][GAT][AG][GC]ACGA[CT][CG][AG][CG][GC][AT][CG][CG][ATC][GT][CG][ATC][GC][CG]T[GC][CA][AGC][GC]C
18	1.50E+00	1.30E+00	8	8	[GA][ACT][GC][AC][TAC][GC][CAG]CG[AC][CT][GC][CGT][ACGTA][CG]G[AT][GC][AC]AG[AC]T[GC][CT][CG][GC][AT][CA][CG][GT][TCG][CG]A[CT]G[CTA][ACT]C[GT][ACT]
19	1.80E+00	6.00E-01	10	6	T[CT]CA[CT]C[CA][GT]C[AG]TCAA
20	4.00E+00	3.70E-02	13	13	ACG[AG][TC][GAC][AT]TCA[TC]
21	5.80E+00	6.60E-05	37	29	[ACT][TA][CG][GC]T[CG][GT][CAG][CT][AGC][AC][CG][GA][GA][CG][GCT][CT][GC][CG][TC]CG[GTC]C[GC][ACT][CG][GC][TC][GC]C[AT][GC][CG][TACGA][GC][CTA]T[CG]
22	1.80E-01	8.20E-02	13	13	[GC]TCGA[AT][CG][AT]
24	2.30E+01	4.50E-02	13	13	TC[AG]T[CG][TAG]AC[GC][AT][GC][AC][TA][GC][GT][TA][CG][AGC][CT]
25	1.40E+00	2.40E-04	5	4	[AT][AT][AG]C[CG][CAG][GC]T[CT][CG][GAT]T[AT][AT][TCG][AG]T[TG][CG][GC][GC][GC][AGT][CA][AT][CT]A[CAG][TAG][CG][GT][CAT]AT
26	1.20E+00	6.00E-02	12	5	[TG]T[GACTA][AG][ACG][ACT][CA][TG][CG][CT][TACGA][ACT][AG]C[AG][ACGTC][AT]A[CT][GT][AGC][AC][ACG][AC][GA][GT][TA][GA][TA]
30	9.80E-01	1.30E-02	42	4	[CT][TG][CT][CG][GT][TA][TAC][CA]GA[TC][AT]CG[CAG][CAG]A[GAT]TGT[AT][CT]CG[AG][AT][AT][AC]
32	7.10E+00	9.10E-01	10	10	C[GA]A[ACG][GT][AC][CAG]CA[CG][TC][CAGTA]CG
34	8.00E+00	5.00E-06	27	14	G[GTA]C[CG][TA][CT]G[TGA]C[CG][GAT][GT][GC][CA][CG]G[TAG][CT]G[GT][GCT]CA[CG]G[AGT][CT]G[AC]

Network	*E*-value$_t$	*E*_value	Cistrons in network	Cistrons with motif	Consensus sequence
35	5.70E+01	2.40E+00	12	11	T[CG][AG]CCG[AG]ACA[CG][GC]T[TC]
36	9.00E+00	5.40E+00	12	9	[GA][AC][CT]G[AT][TG][CGT]A[CA]G[ACT][CG][AT][TG]C
41	2.30E+01	1.10E+01	7	5	[GT][GT][ACGTC][CAG]G[ACT][CGA][CG][TA][CAGTA]G[GTC][CG][GT][GT][CA][GAC][AGT][CG]GA[CGT][CGT][ACT][GAC]G[CAG][AGT][CG][GAT][CG]G[ACG][CA][GC][TG][TC][GA][AT][CG][GA][GAC][CT][CT][AT][GC][CG][AC][GA]
42	5.20E+00	5.10E-01	13	13	G[TC]CGAC[CG][AC]G[CG]T[CG][TG][TA][CT]
44	7.10E+01	2.30E+00	13	5	[CT]G[AT][TA][CGT]G[ATG]C[GA][TA][CG][AGT][AT][GCT][GA][TAG][CAG][GCT][AT][CG][AT][TAG][CT][CG]A[GTC]C
45	3.70E+00	6.20E-03	8	8	[CG]ACCG[CAG]C[AG][AT][GT]CTC[GCT][CA]C[GA][TC][GC][ACG][AC]
46	2.90E+01	1.20E-01	17	8	[TC][CG][AT][CGT][GC]TC[GC]G[TG][GC]ACG[ACG][AT]C[AG][CG]GT
47	7.90E+01	2.80E+00	10	2	GAAA[CG]T[AT][GT]CAGT[GT][CT][CT]AA[CT]TT[AC][AT]ATG[GT][AT][CT][CG]G[AC]T[GT][AT][CT]A[AG]GA[CT][GT]T
49	4.70E+00	6.70E-01	24	24	[GC][GA][TCG]G[GA]T[CG]A[CG][CG][GT]
51	3.40E+01	1.70E+00	4	4	T[CG]G[ACGTA]C[AC][CG][CG][AC][CT][CG][CT]T[CG][CAG][CA][CAT][GAT][CAG][GA][TA][AT][CT][CG][TCG]CG[GA][GAC][AT][TA][CG][TG][TA][CG]G
55	6.70E+00	2.30E-03	13	12	ACGAC[GC][TA][CT]CT[CGT]G[TA][CT]
56	7.10E-01	2.20E-05	54	54	[TA][CG][GC][TA]CG[TA][CGT][CG][ACG]CG[AG][CG][GT][TA]C[GT][TC]CG
57	1.50E+01	8.40E-01	8	7	[AC][GA][CT]GTC[CA]AC[CG]G[AC]TC[GCT][CT][CT][GC][CAG][CT][GC][AT]TC[TG]CCT
58	2.90E+00	1.30E-02	15	9	[TGA]C[CG][ATC][GC][GC]A[CG][AG][ACT][CG][GA][TG][CG][TCA][TA][CGT][GA][AGT][CG][GA][CTA]C[AC][CT][CG][GT][CT][CAT][CG][ACG][GC]G[CA][CGA]C[GT]GA[CT]C
59	6.70E+00	5.70E+00	11	11	[TA]C[GC][AG][CGT]G[AT]CG[AT]CGA[GAT]G
60	7.60E+00	2.40E-01	22	10	[TC][CG][GA][CT][CG][TAG]TC[GC][GAT][CG][AC][AC][GC][ATC][TGC]C[GA][TAG]C[GCA][GCA][GC][ACG][GAT]C[AC][CT]G
62	1.80E-01	5.90E-02	24	8	T[TG][CGT]G[TC][GC][TG][CG]C[AG]T[GC][ACT][TC][GC][TC]G[ACG]C[AT][ACG][TC][TA]TGA
63	1.80E+01	3.80E-01	5	4	[AG][ACT]T[AT][GT][TAG][GT][CT][AG]TG[AT][CGT][ACG][CT][GC][GAC][AT]T[CG][TA]
64	4.20E+01	1.60E+01	9	2	T[AT][CT]AC[AG][CT]TGC[GT]TGT[AT][AT]AT[AG]CGC[AG][GT]T[AT][CG][AG]TA[AG]TG
65	5.00E+00	1.10E-01	10	4	TT[AG][TA]T[GC][CA][CAG][CT][AG]T[GT][AT][GT][GT]T
66	8.90E+01	1.90E+01	11	6	[TA]GTCC[GC][CG][CG][TG][GT][AC][AC]GT[TA][CT]C[CT][TC][AT][GA]C[AT][CG]A[GA][ACT][CA][CA][ACG][AC][CT][AC]T[TA]
68	2.60E+00	2.30E-05	26	14	[GC][TG][CG]G[ACGTC][GC][CAG][AT][GCT][CG][TAG][CG][CA][ATG][CG][GA][AC][GC][CG][AT][CG]G[TG]C[GC][CA][GC][CG][GCT][GCA][CG][AT][CG][ACG][TC]C[GCA][CTG]C[GAC][AC][AG]C[TA][CG][CG][CGT][CG]G[TG]
69	7.30E+00	2.40E+00	21	18	CG[TA]C[GC][AG][CGT][CTG][CG][CG][GC][TG][CG]G[CG][GT]G[CAT]C[CG][ACT][TCG][CGT]T[GC][CT]TC[GC]
70	3.70E+00	6.00E-02	16	10	[AG][TC][CG][GAC][TCG][GA][CA][CGT][CGT][GT][CT][CGT][AT][TC][GC][ACG][TG][CT][GA][CAT][AGT][AGT][TA][CGA][GTA]C[CGT][ACT][AG]C[AG]A[CT]
72	1.10E+01	7.40E+00	9	6	TG[CG][CA][AT][TG][CG][TA]CG[CA]CA[ACT]CG[CT]CG[AT][AT]
73	1.10E+01	7.10E-02	17	4	T[GT][AT][AG][TAC][GT][ACGTA]A[TCG][GC][ACGTA][TAC][GA]T[GT][AC][AG]C[TA][ACT][GT][CAG][CGT][AT][ACT][CAG][CAT][GC][GAC][AT][AG][AG]T[CG][CAG][GAT][CGT]T[GAC][GAT][AC][GT][ACT][TC][TA][CGT][CT]A
74	4.60E-01	2.60E-05	21	13	[TA]G[AG]C[GC][TAG][TC][CG]G[GCT][CG]A[CTA][GC][GT][CTA][GC][CAG][AT][GC]GTC[AG]T
75	3.50E+00	2.10E+00	6	6	[AC][GA][AC]G[CG][CT][CG]TAC[AG]C[ACG][AG][CG][AC]GC[TA][GT][CT]CG[AT][CA][GT][GA]GATCT[TC][GT]CC
76	1.20E+00	4.30E-01	10	5	[TC][GT][AT][TA]CTT[TC][GA][CAT][CGA]G[TC][ACG][TAC][GA][TAC][TA]
78	5.70E+00	2.70E+00	8	6	A[AC][CG]A[GA]CG[TA]C[CT][AT]C[CG]TG[GA]A[GT]G[TG]
79	3.80E+01	7.10E+00	9	5	[AT][CGT][TG]T[TC][GCT][AT][CA][TG]AG[GTC][TA]G

Network	*E*-value$_t$	*E*_value	Cistrons in network	Cistrons with motif	Consensus sequence
80	3.90E+01	7.30E+00	25	4	[AT]G[CG][AT][TA][CAG]A[AG][CA]A[AG][CAG][AC]TC[TA][AG][GC][TC][AT]
81	1.10E+02	3.70E+00	7	7	[AT]CG[GT]G[GC][TA]GACCG[AC]G
84	1.50E+00	1.60E-03	15	11	TC[CG][AT][GA][CG]TC[GC]T[CG][AGC][AT]T
85	2.20E+01	1.30E-05	27	14	[CG][CA]T[GC][CG]CC[GA][TAG][GC][GAC][TA]C[GCT][TC][CG]G[ACGTG][CG][CA][TA]G[GC]T[GC][GTACA][TC]
86	1.40E+00	5.20E-04	18	18	[TA][CG]GAG[CGT]TG[CG]T[GC][GT][AGT]CC
87	1.00E+02	1.10E+01	11	11	TCGT[CG][CA]A[CG]GTC
88	3.80E-01	1.10E-01	32	16	G[ACG][CA]G[AG][TC][CG]T[CG]CT[CT][CG]A[GT][CG][GA]C[CT][TA]C
89	4.30E+00	3.70E-01	9	8	[AG]G[AC]T[CT]C[GA][AG][ACG][CT]G[AT][CG][AG][AT]C[AT][CG][CGT][AT]
91	2.70E+01	6.00E+00	9	5	G[TC]T[CT][GA][AT][CGT][GA][TC][CA]AT[GC][GT][CG][CA]G
92	1.30E+01	3.20E+00	11	8	[AT]C[CT]T[CG]C[GT]C[AG]AC[GA][CAT][CT][CT][TC]CG[CA]CG
93	8.50E+00	3.30E-01	11	5	T[CA]G[TA][CT]G[CTG][CG][GA]A[CGT]G[CGT][CTA][GC][GACTA][CAGTA][GT][TA][CAGTA]G[ACT][CA][GT][ACG][CA]G[GC][CAGTA][CGT][GA][GT][CA][TA][GC]G[AGT][CA][GC][AGT]G[GCT][TCG][CAG][CG][TCG][CGT]G
95	2.90E+01	8.20E+00	12	8	[AT]CCGC[GC][CG]G[TC]ACC[CG][GC][GA][AT][ACG][GA][TC]
96	8.50E+00	7.60E-02	25	18	[AC][AT]G[TA]TC[ACG][AT][GCA][AG][TCA]C[GT][TCA][CG][CT][TGC][CG][GA][CT][GC][GA][AT][GCA][TCG][GT][CG]G
97	1.30E+01	1.60E-01	14	14	[GT]T[CG][AG][TA][CG]G[AT][CG][AT]T
98	3.70E+00	9.10E-01	8	5	[GA][TA][CA][GA][TA][GA][GC][TA]T[CG][TA][ATC][CG][AGT][GA][GAT]T[GTC][GAC][CAG][TC][GCT][AC][GTA][TCG][GA][TAG]
99	1.30E+00	4.50E-05	16	8	CG[TA][GC][GA][TA][CG][GT][TA]C[CG]AG[ACG]T[CG][CAG]TC[GC]T[CGT]G[AT][GTC][CG][AC][CG][CG]TGG[AG][CA]G[CT]CG[CTA][GCT][CT]
100	3.00E+01	4.40E-02	14	4	[AC][AT][AGT][TC][TG][ACT][GT][CGT][CT][ACT][CAT][CA][GA][AC][TA][CT][AGT][TCG][TC]A[ACG][GT][GCT][AT]AA[TG][CG][CG][AT][CA][ACGTA][GA][TCG][AC][AG][CAG][AT][GAC][AC][AT]
101	4.10E+01	1.00E-02	14	4	[AG][CGT]G[AT]A[TC][TG]A[GCT]T[AC][AC][TA][CGT][TAC][CG][CAT][TA][CGT][AT][AG][ACGTA][CAG]A[TA][AC][TA]
102	2.90E+00	1.00E+00	6	3	[TA][TG][TG][AGT]A[AC][TA][CGT][AT][AC]T[GA][CGT][AC][AT][TC][GT]CC[AG][GA][AT][TC][CG]AGTT
103	3.80E+01	7.40E-04	13	13	C[GT]G[CA][AGC]CGGT[CG]A
104	1.20E+01	9.70E+00	18	5	[TA][AGT][CGA][TA][CT][GA][AT][TA][GTA][ATG]T[GC][AC][GC][CGT][AGT][TA][TC]A[TACGA]G
109	2.00E+00	1.90E+00	5	5	[AC][TA][TC]TC[TG][TC]
110	2.50E+00	8.30E-03	12	12	[GA][TC]G[AG][TC][CG][AT][CA]G[AC][AT][CG]T[ACGTA]C
113	9.20E+00	2.50E-02	26	26	[TG][CG][CG]C[GCA]G[GA]CG[TAC][CG][GC][TG]CG[GT][AC][CG]A[CG][CG]T[CTG][GC][TA]CG[CA][CG]G[AT][GC][GTC][TCA][GC][CG][AG][CG][GC][GA][GC][GCT][AC]C[GC][AGT][CG]
115	3.00E+00	1.90E-03	13	5	[TG][ACG][AT][CT][AC][GA][GC][GTC][CGT]A[GAT][TC][GCT][CAT][TC][ACT][CG][CT][TC][GA][TG][AG][CTG][TG][CA][ACT][CGT][AT][AT][CAT][AG][CA][CA][GC][ACG][TG][GC][ACT]T[CA]
118	1.10E+01	5.80E+00	6	6	CGGC[AG]T[CG]TCGT
121	5.60E+01	3.10E+01	13	10	T[GC][GC]AC[CG]G[CGT][AGT][CGT][CG][CA][GA][GCT][TG][CT][CG]G[TGA][GCA][GT]
123	1.20E+01	1.00E+01	12	12	[AC][GAC]CT[GC][GC]TCC[AT]
124	8.40E-01	6.50E-03	17	13	[TG][CG][CGT][AC]C[GC]A[TG][CG][TC][CT]G[TG][CTG][GC]ACG[CG][ATC][GC][GC]TC
125	6.90E+01	7.10E+00	14	9	[AT][GC][CG]A[CG]GTCGTC
126	2.80E+01	1.80E-01	6	4	T[GT][TG][GC][TA][GA][ACGTA][GC][GAC][GC][AG][TCG][GCT][TG][AG][TA]C[GAC][GA][CAG][TCG][AT][AT][GCT][GA][GT][AT][GT][GC][AC][ACGTA][TC][AG][TCG][ACG][TA][CAG][GT][GA][GAT][CT][CG]C[AT][CT][GA][AC]A

Network	*E*-value$_t$	*E*_value	Cistrons in network	Cistrons with motif	Consensus sequence
127	5.70E-02	3.50E-02	11	7	T[CT][GC][TG]T[GC][TA][GC][GC]ACG[CT][CG]GACG[GT]
128	2.40E+01	1.80E-01	13	10	GACGG[TAC]GAA
129	3.00E+00	1.90E-03	16	14	[TC][TA][CT][ATC]T[GC][AC][ATG]C[GA][GC][ACG]AA[CG][CTA][CG][TAG]
131	1.90E+00	1.70E-02	8	8	[TG]G[CTA][CG][GAC][AT][TC][GC][AG][CT]GAT[CG][AC][AGT][CT][GC][CG][GT][GC][TA][GT][TAG][CGA][GC][TG][TG][CT][ACG][TG][GA][CGT][GCT][CGT][GC]G[CA][TG][CG]A
133	4.50E+00	1.20E+00	10	9	CGACGAC[AC]
135	2.30E+01	1.30E+00	11	4	[AT]T[GC][TA][GCT]AT[GT][CT][GT][TG][CG][AT][TAC][TCG][GAT][AC][CG]A[AT]
137	2.10E+01	1.90E+01	2	2	T[AC][CT][GT]CGTAG[CG][AG]C[AG][AG][CT]CGAG[AT][AT]TTC[GT][CT]AC[CG][CG][AT]A[AT][AC][CG][CG][GT]C[GT][GT]A[AG]T[AT]T[CT]G[CT]T
138	2.80E+00	4.80E-02	11	9	[GA][CT]C[GC]C[CG][GA]T[GA][GA]TC[GA][ATG][CG][GT][GC]C[GA][TA]
139	1.20E+00	3.40E-04	9	6	[AT]TGGCG[AT]A[GC][TG][TC][CG][AT][CG][GC][TA]CC[CG][TC][GC]A[AC]G[GT]CG[GA][CT]C[AT]G[GC][AT]C[GC][AG][TC][GC][TC][CA]G[AG]
141	2.10E+01	7.70E+00	6	5	[AT][GC][AT][TAC][CTA][CGT]GA[TACGA][GACTA][AG][TG][CGT][AC]T[CGT][TG][GC][AT][GC]C
142	1.90E-01	1.40E-01	8	4	[TA][AC][AT][TAC][ACT][AC][CAT][GA]A[CG][TAC][GCT][CGT][AT][TA][CT][AC]C[GC][GAT][AG][GAC][AG][GA][CAT][GAT]T[TA]C[AC][ACT][CAT][ACT][AC][CG]TG
143	4.00E+01	1.90E+01	11	8	[AT][CG]G[AG]C[CG][ACGTA][CG][CA][AT][GC][AG][CAT]CC[TC][CG]G[AT]C[GC][AC][CG]GT[CA]CTG
144	1.30E-02	6.80E-03	10	7	[AT][GC]T[AGT]T[TC][GT][AT][CAG][CT]GT[CT][AG]T[AT][CG][AG]T[CG]T
145	4.60E-01	2.70E-01	10	4	[AT][GCT][TG][GCT]G[TG][CA][TC][TAG][GT][AT][CT]C[GT]C[ACGTA]A[GA][AT][CT]G[GA][AC][AT][CT][AT][CT][TA]
146	6.30E+01	3.40E+01	6	4	[AT][CT][TG][CT]GT[CT][TA][TA]G[AG]
148	8.10E+00	6.50E+00	7	4	G[AG][GA]T[GA][CA][AG][CGT]TT[CT][AT]T
150	1.20E+02	5.60E-04	11	8	[CT]GC[AG]C[AGC][CG][CG]G[GC][CG]GT[GC][GC]TCG[TA][CA][GC]T[CT][GT][GAC][CG][GC]AC[GC][AT][CG][CGT]A
151	3.00E+00	4.80E-09	9	4	[AT][TC]TG[CA][AC][TG][AG][AG][AGT][CT][AG]TGC[AG][TG][ACT][GAC][TAG][CGT][AC][TC]G[CAT]A[TA]A[CG]TC[AT]T[GCT][CG][AT]
152	3.60E-01	3.40E-02	21	16	G[AT][CA][GC]AG[CG][AT]CG[TG][CA][GC][TG][CT]
154	1.50E+00	7.30E-02	9	5	[TC][GT][CT][CGT][TC][ACT][CGT][GT]T[TG][CTG][TA]C[TA][CT]G[GT]T
157	7.10E+00	6.60E+00	8	8	[TG][GC]CT[GT]TCGA[TA]
159	2.10E+00	9.80E-01	16	7	GGAAGT[CG]C[ACT]G[ACG][TG]C[TA][TC][CG]GG[CG][GA]C[AC]GG[CAT][TG][AG]C
160	2.90E+01	9.70E-08	28	8	[AG][GT]TT[GA][AC][GC]TAC[CG][CGT]G[TA]AC[TC][CG][AT][TAG]
161	2.00E+01	9.20E+00	16	9	CG[ATC]CCT[GC][CGA][TC]CT[TAG]CGC[GC][AG]T
165	2.30E-01	2.80E-03	17	7	[GA]A[CG]A[CG][GC]G[AT]GGT[GC][AG][TG]G[AGT]T[GC]GT[GC]A[CAT][CG][CG][AT]G[CAG]T[GC]G[AT][CG][GC][AC]C[ACG]TC[GCT][TA]
166	6.90E+00	2.80E+00	6	4	A[TA]C[GT][TA]T[CA][TA][GAC]T
167	5.00E+01	1.70E+01	11	7	GTCC[TA][CA]CGT[CGT][AG]AGG
169	6.20E-02	6.00E-02	10	7	[TC]C[TA][GA][GC][AC][AC]T[ACG][GT][TG]CC[AGT][TA][CAT][CG][CAG][GT][ACG]C[CG]A[GT]T[GC][AG]A[TG][GC]A[GCT][TG]G[CG][AT]CC[GT][AT][CA][CG][GA]G[GC][GA][CAT]G[GT]
170	5.60E-02	1.30E-03	7	7	[AG]CT[CT]T[AT]TC[TG]T[CG]A
171	5.40E+00	8.40E-04	13	13	[GC]A[GAC]C[AG][CA][GC]A[CGT][CG][ACG][GA]G[AT]CG[GT]CG[AGT][AC][CG][AGT]C[GA][TG][GC][GC][TCA]

Network	*E*-value$_t$	*E*_value	Cistrons in network	Cistrons with motif	Consensus sequence
172	1.80E+01	6.90E+00	9	8	[AC][CG][AG][CT]CG[CA][CG]GAG[GC][AT][CG][GTC][TC][CG][GA]AG
173	3.50E+01	4.30E-01	15	5	[CA]G[AGC][CG][GC][GAT][CAT][GC]G[CT][GC][AG][GC][CGA][GC][TA]G[CG][TG][GC]G[CT]G[GC][TCG]G[GC][TC][GC][AC][CT][GC][GT][TC]G[CG][TG][GC][GC][TA]G
176	5.80E+00	4.70E+00	30	4	[TA][TAG][CAT][AT][AT][AC][CGT][AT]C[TAC][GA]C[TA][CT]T[GCT]T[GC][AG][ACT][AT][AT][CG][AG][TC][CAG][TAC][AGT]G[AC][GA][CT][AGT][GT][CAG]G[GT][AC]
177	3.70E-01	3.20E-01	19	4	[AT]AAG[TAC][ACT][ACGTA][AT][AC][ACT][TC][GT][TC][GA]C[AC][CA][CAG][ACG][TCG][AC][TC][ACG]T[TC][CG][TCG][CG][GAC][AC][ACGTA][TC][CT][GCT][AT][TA][CG][AT][CGT][AT]C
183	3.40E+01	9.90E-05	20	18	G[TA]G[ACG][TA][CG]G[AT][CG][CG][AT][GCT][CG][TA]C[CG][TG][CG][GT][TCG][CG][GC]A[CG][CG][AG][CG][GCA][TAC]
184	4.40E+00	2.60E+00	12	5	[CT][TG]G[GA][TAG]C[GC][TCG][GC]A[CT]GG[ACGTC]C[GAC][ATC][CGT][GCT][AGC][CAG][AC]TC[GT][AT]C[GAC][GAT][ACG]G[TC]G[GA]
185	2.10E+01	3.00E+00	8	5	[TG][CG][AT][AG][CAG][AGC]A[TC][CG][CG][TAG][GT][GACTA][ACGTC]C[GA][ACT][GAC][GAT][GAC]T[CT]ACC[ATC][ACGTC][CA]G
186	6.00E+01	4.90E-02	13	7	CG[TG]C[GC][TAG][CGT]G[TC]CG[AGT]C[GC][AG]T[CG][CAT][CT][CG]T[CT]CTCG[AT]C[GC][GT][TC][CG]T[TC][CG][AG][AC][GC]G
189	1.50E+01	4.50E-01	10	5	[AT][CGT][AGT][TCG][CG][AGC][TAC][GC][TA][TAG]C[ACGTC][ATG][CT][GAC][GTC][CAG][AG][GT][TC]G[CA][CGA][GC][ACG]G[GA]T[GC][AG][TA][GC][CGT][AC][CG][AC][AT][CG][CGA][TA][CT]
192	1.40E+01	4.40E+00	13	13	G[GC][TA][CA][CG]TCGT[CA][GC]
193	8.20E+00	4.00E-03	18	6	[TA][TC][CG][TG][CT][TA]C[CT][AG][AT]G[GA][AC][AC]A[AC][CA][ACG][AC][GT]T[CA][CG][AC][CA][CA][GA]A[GC]G[AC]TCAC[TC][GC]GTC[CA]AT[CG][CA]AA[AGT][AG]
195	1.20E+01	1.40E-02	12	8	CA[GC][CT][GTA]GG[CT][CA][GC]AGC[AT]GG[AT]C[CG][AT]C[GC][TA][GCT][GT][AT]GC[GCT]C[CG][TG][GA]GC[GAC][GA][AT]
200	7.00E+01	2.60E+01	10	5	[GT]C[CGT][GAC][TAC][GT][AT]CC[TAG][CA]C[CGT][GC][GT][GT][CGT][CTA]TC[CT][GC][CAGTA][CA][AG][TA][GC][TA][CG][CG][ACGTC][CA][CGA][AG][CAGTA][CGA][AT][ATC]
201	7.60E+01	2.00E-02	22	10	C[GA][TC][CG][GA][AC][CT][GC][AGT][CGATA][GCT][ATCGC]C[GC]G[CT]GA[GT][CGT][TAG][CG][TACGA][TC]CG[AT]C[CG]
202	1.20E+02	1.50E+01	25	21	[ATG][GC][GA][CAT][GC]G[ATG][CA]CT[GC]GA[CG]G[TA]CG[GT]C
203	2.40E+02	6.20E+00	12	11	G[CG][GC][GA][AC][CG][GC][GT][GC][GAC][CA]GC[GA][CG][CG][CG][CG]TA[CG][GC][CG][CG]G[GC][CG][GC][AG]GG[GAT][GC][GA]CG[CAG]AG[GC]TCG[AC]CGC[CA]C[TCG]
205	4.50E+01	1.00E+01	5	5	AT[CG][AGT][TA]CA[CT][CG][ATG][AG]
206	8.30E+00	7.10E-02	24	18	TC[GC][GT][CT][GC][CA][AT]G[AT][CA][CG][GA][CG][GC][GT][TCA][GC][TGA]C[GC]G[CA]G[ACG][CT]G[AT][CT]
207	9.00E+00	2.60E+00	7	4	A[AT]C[TAG][GCT]A[AC][TCG][GAC]A[AC][ACG][AGT]T[CG][AG][TAG]T[CT][ACG]C[CAT][CT][TA]
208	3.30E+00	8.40E-01	13	9	[AG][TC]G[TA]T[CG]A[CG]GG[TA]G[CGT][CT]G
209	5.80E+01	3.60E-02	17	5	[AC][GCT]A[ATC][GTC][GC][GTC][GA][AC][CTG][GT][AG][ACT][CT][CAGTA][CT][GC][CTA][CG][GTA]T[GCT][TC][GT][CTG][CGT][CGT][CTA][CGT][GTA][TG][CGT][TG][GT][ATC]TC[CT][ACG][ACT][GACTA][CAGTA][TA][AGT]TC[GC][AG][TC]
211	3.70E+01	2.00E+00	18	9	A[AG][CG][AT]C[AG]TGAA[CA]
212	8.60E+01	6.00E-03	17	8	G[AT][CT]G[CA]CC[AT]GG[AC]CG[GA]CG[CA][GC]CG[CG][GA][AG][GT][GAC]CCG[AC][GT][GC][AC][CG][GC][GA][CT][GA][CA][GC]G[AC]CG[TAG]CG[ACT][CGT][GC][TC]
213	1.70E+01	1.40E+00	12	7	AA[CG][GT][GC]T[GC]A[CAG]G[AT]AA[TA][AC]
214	1.80E+01	2.10E-04	18	14	[CG][AT][GC][CG][TG]C[CG][GT][CT][CG][AGT][TCG][GC][GAT][CT][CG][AT][TC]G[CG][CA][CG][AG][GC][CG][GAT][CG]G[AG][CGT][GA][TAC][CAGTA][CG][GAT][CTG][GC][AG][GTC][GC][AC][TCA][GC]C[CG][GC][TA][CAT][CG][AGT]
219	1.40E+01	8.10E+00	8	8	[CG]A[CG]G[AG]CGA[CG]G[AT]C[CG]T

Network	*E*-value$_t$	*E*_value	Cistrons in network	Cistrons with motif	Consensus sequence
220	1.60E+01	1.00E-04	21	12	GCCG[AG]G[AG]T[CG][GT][CA][CG]GAC[GAC][TA][GC][GC][TA][CG][GA][TA][CGT][GA][GT]C[GC][CGA][GT]GCC[GA][CT][CG]G[TC][CG]G[TA][CG][GC]T[GC][GC]
221	3.80E+00	2.20E+00	11	6	T[CT]TC[CG]ATCAT[GC]G
225	4.00E+01	3.70E+00	13	10	[AG][ATCGC]G[TC][CA]C[ACG][GT][CGA][GA][CT][CG][TAG]C[GC]TC[CG][TG][CG][GC]
227	1.60E+01	1.30E-11	36	28	[GC][CA][CG][GC][GC][CT]C[AGC][CG]G[TAG][GC][CG][GA][CA][CG][GA][GAT][CG][AC][CAT][GC][AGT]C[CG][TG][CG]G[TAGCC][CAG][CG]G[CT]C[TGA]CG[GT]C[CG][GCT]C[GC][AG][CT][GC][ATC][GTC][CG][CAG]
228	5.20E-02	3.10E-02	27	13	T[CG][AG][AC]C[CTG][TG][CG][TG][TG]C[AGT][TG][CG][GCT]CC[ACG][ACG]C[GC][AT]CA[AC]C[CAG][TG][CT][GCA][TAG][CG][ACG][AT][CG][GC][TGA]C
230	8.00E+00	5.40E-01	24	5	[TA]G[CGT][TG][GA]A[TG][CAG][GC][CAT][CGT][GA][CAT][GC][GTC][TA][GAT][GC]C[CT][GAC][CAGTA][GTC][TA][TA][GC][TA][TA][CT]G[AGC]G[ACGTC][TAC][GT][GCT]G[CA][ATG]
233	7.00E-01	1.10E-01	9	9	[AG]T[CG][AG][CA][CG][CG][TG][GC][AT][TCA][GC][TCA][GCT]C[GT][GC][ACT]CA[TGC][TC][CT][GC]G[TAG][CT]GA[CG][GC][AT][AC]
234	8.10E+01	3.30E+01	7	4	[AT]C[CG][GCT][TA][CT]TC[CT]T[CT][AG]C[CG][ACT][TA][CAG][CG]A
235	1.20E+01	1.70E+00	9	7	CGGAG[GA][AT]CGA[CG][GC][AT][GC]A
238	9.10E+01	1.70E+01	7	5	A[CG][AC][TA][GAC][GC][AC][CG][GA][AG]G[GC][TCG][GAC][TC]TG[AG]
239	3.50E-01	4.50E-02	11	11	[AC]CTC[AG]AC[CG]G[CT]G[ACT][CG][GA]A
240	2.20E+01	1.10E-02	19	7	[TG]TGA[GA]CA[AC][CG][TAC]C[GC][GC][CG][CT][GC][TCG][CG][GC]C[GT]T[GC]CGTA[CT][GAC]G[AG][TC][CGT]GG
244	3.90E+01	6.40E-01	11	5	[AC][AG]A[CT][CAT][AG][AG][AT][TAC][TA][GAT][CGT][GACTA][CA][AT][CAT][GACTA][AT][CTG][TA][ACG][ACT][AG][TAG]
246	1.40E+00	1.20E+00	10	10	GTTCA[TCA][CG][TG][CT]
247	3.30E+01	8.80E-07	19	18	[AG]CG[GTA]C[AG][AT]C[CG][AGC]C[GC][TA][CG][CG][TAG]C[CG][ACT]C[CG][TG]C[GA][GAT]C[GAC]AG
248	1.10E+00	3.40E-01	24	8	[CT][ATG][GT][GC][TG][TA][GCT][GT]C[CG][GT][TG]C[GA]A[CG][GAT][ACG]G[ACG]A[CG][AGT][TAC]C[AT]T[CG][CG][CG][CAG][AC][AT][GA][GT][GAT][CG]TT[GC][AG][TC]
249	8.30E-01	1.40E-20	51	44	G[ACG][CG][GC][TAGCC][CG][GC][TA][GC][ACG]T[CG][GA][ACT][CG][CG][GTA][GC][GA][TAG]C[GAC][TG][CG][GC]C[GC][GC][ATC][CG][GC][TAG][CG][CG][CAG][GC]G[TC][CG][CG][TA][CG][GC][ACG][CG][GC][TC][CG]G[ATCGC]
250	3.80E+00	5.90E-01	9	9	[TC][GC]TTCGT[GA]C[GC][CAT][CT][CG]
252	6.40E-01	6.00E-02	15	10	C[GA][TG][GA][CT]G[AG][AGC][TAG]AG[AG][AGT][TCA][CG][AT][TCG][TAC][GCT][CAG][GT][GAC][AGT][TCA][AC][CT]
253	2.30E-02	5.60E-04	10	4	[TC][GCT][GA][AT][AT][AC][GA]T[CT][CT][AC][GT][CGT][CG][ACT][CAG][CG][GAT][AT][CAG][CA][CT][CA][GAT]A[AC][ACGTA][GC][TG][AC][CG][AT]T[AG][AC][CT][GT][AG][TCG][CG][TCG][TG][AC][AC][ACG][GT][TAC][AG]A[AG]
254	8.60E+01	7.30E-06	30	20	[GTC][TA]C[GC][AC][CG][GC][AC]C[GC][ACTGG]CG[CGT]C[CTG][AT][CG][GC][GT][CG][GA][AG][CG]C[GT]G[CT][TGACA]C[GC][CGT][CG][GC][TCG][CG]G[ATG][GC]GA
255	6.90E+00	6.60E+00	12	11	T[GC][CT][GT]G[GA][AC][GC]A[TC][CT][TG]CCT
257	1.80E+01	1.90E-02	9	6	T[GT][TC][CT][GC]G[TG][AGT][TG][CA]GTCG[AC]CAT[AC]A[ACG]G[AGT][GC][AC]C[GT]TC
259	3.60E+01	4.60E-02	13	6	[AT]C[GA][AC][AG]GTC[CG]G[CG]G[CA][AC][GC][TA][TG][AG][AT][TG][CA]TG[AT][GC]T[TA]C[ACG][TG][GC][CGT][CT]A[TC]GG[GT]C[AG][AT][AT]CGG[TC][ACT]
262	4.00E+01	1.40E-01	15	7	CAC[CG]CC[GCT]G[CAT][TC][TG][GT][TC][CG][GT][CT]C[GT][AC][GAC][TCG][GC]T[GCT][GC]CTAC[TC][TAC][TC][CG][TA][CA]TT[TC][AG]C
263	1.00E+01	1.70E-02	12	11	G[GT][CA][CG][AG][CG]CAG[GT][TAC][CG][CG][AC][CG][GC]A[AC][GC][TAG]C[GC][AG][TC][GC][GCT][CT][GC][CA][GC][CG][GAT]CGA[TA]G[AT]T[CG][AC][GTA]C[GA]C[CG][ACT]C[GC][AC]
265	3.20E+00	7.30E-02	16	9	[AG]GTGGT[TA][GC][AG][AT]

Network	*E*-value$_t$	*E*_value	Cistrons in network	Cistrons with motif	Consensus sequence
266	4.00E+00	5.70E-04	7	4	[AC][TA][CG][CT][GT][TAG][CAT][CT][GAT][CT][GT][CT][CA][GA][AT][CG]A[AG][AT][CAT][TG][GAC][CT][GT][CG][ACG][GAC][ACT][CT]T[ACT][TG][GT]T[CT][TA][AT][GA][TG][GCT][CAT][GAT][CAG][AG][AC][GC][GCT]A[AG]
268	4.10E-02	2.90E-04	13	11	CATC[TA]G[GCT][AC][GC]AA[ATC]TT[CG]
269	1.00E+01	4.90E+00	8	7	GAC[GT][TA]GA[TC][CT][TA]T
270	1.30E-01	3.90E-10	19	3	A[TG]AT[GT][GC]A[AT][AC]A[TC]C[TA][AC]GT[CA]C[AT]C[TA]A[GA][CA]T[CT][TA][AT][AC][CT][CG]TAG[TA][TA][AC]A[CG][TG]A[GT][CT]TT[CT][AG]C[TC]
272	3.40E+00	6.00E-01	10	4	[AT][CT][GT][TG][TA][CA][AG][CAG][TG][ACGTA][GA][TCG]C[AC][GT]C[AG][AT][GT][AG][CAT][CT][CG][TG][GC]A[GT][AG][GC][TA][GT][CG][GAC][CGT]C[GCT][AG]GA
274	7.20E+00	1.60E-01	9	5	[AGT]T[TA][CG][AT]T[TA][CT]G[ACT][TA][GT]
276	5.70E+01	6.90E+00	10	4	T[GCT][AT][ACGTA][ACGTA][CG][TCG][GAT][GA][AC]C[AC][TAC]G[GCT]A[AG][ACG]T[GC][AG][AT][CG][CG][AT][AT][CG][CA][CGT][GC][GT][CT][ACGTA][AC][GAC][TC]T[AG][AC]
278	3.70E+01	3.10E-01	11	10	[GT][TCG][CG][CAG][AC]G[CT][CAT]CGA[CA]G[AGT]CG[AT][CG][CG][TC][CG][CAT][CTAGA][GAC][AGC][ACG][CT][GA][TG]
279	5.10E+00	4.90E-02	6	6	G[AT][GC]C[TA]GGCCGA[GC][GA][GC]GC[TA]C[GC]A
282	7.50E+00	7.10E+00	14	5	[CT][AG][GACTA][AC]A[GTA]CG[CGT][TA][TG]C[CA][ACGTC][CAT][TA][TC][CG][TC][TAC]C[GACTA][TA][GCT][TAC][AG]
283	2.40E+01	1.50E-03	25	21	AG[CG][TC][GT]GTCG[AT]CA[CT]
285	3.60E+01	2.30E+01	13	9	[AG][CG][CG]C[TC][GC][CAT][CTA][GC][TA][CTG]C[GC][TC]G[AGC][AT][CTG][ACG][CA][CT]C[TGA][CG]G[TC]C[AGC]
286	1.10E+02	5.50E+01	8	2	T[AT][CT]T[CT]CAA[GT][AG][AC][AC][CT]ATG[AG]TTC[CT]CAGG[AT][AT]T
289	3.20E-01	2.30E-04	39	33	C[GCA]T[CG][CG]TCG[GTC]CG[ACT][CG][CA][GA][GC][CTG]TCG[AGC][CA][GC][AGTCC][CT][GC][GCA]C[GC]
290	1.10E+00	3.90E-02	8	5	[CA][CAT][TA][CTG][CTG][GAC][CG]T[GA][TA][CAT][GT]A[AGT][ACG][ACT][GCT][CAT][TG][CG][AGT][ATG][GC][GACTA][ACGTC][ACGTC][TACGA][CAGTA][ACG][AG]T[CT][GAC][ACT]T[CAT][GTC][TA][GT]
291	6.20E-01	1.80E-01	9	5	T[CGT][TCG][TC][CG][TAG][TA][CTG][AGT][AC]G[GC][TA][ACGTC][GC][GT][GCT][AGC][AG][CT]CG[AT][CT][TA][GT][GAT][CG][TC]
295	1.90E+00	3.30E-02	15	10	[TC]CG[GA][TC][GT][GCT][CT]G[AC][CG][GCT][TAG]C[AGT][CAT]C[GC][CG][CAG][GC][AC]CG[ACG]C[CT][TAC][CG][GC][CG][CG]G[AGC][CG][GC][AG][CG][GAT][TAG]C
299	2.40E-01	2.20E-02	13	4	[AG][AC][TG][ACT][AC][AG][CAG][AT][AC]A[GAC]A[CAG][GT][TA][TA][CT][CA][GC][GA]A[AC][AGT][ACGTA][GC]GT[TA]
304	6.00E+01	4.30E-02	17	12	[CT][GC][GCT][TC]GA[TC]C[TG]C[CG][GT][GCT][GC][AT]TC[CG]C[GC][GA][AT]GG[GAT]C[ACGTA][TG][GC][AG][CG][GC][GC][ACT]G[GA][CT]G[GTA][CTA][CG][AG][TG]CGG
305	5.20E+01	4.40E+01	4	3	AT[ACT]GT[CT]CT[GT][GC]T
308	1.90E-01	3.80E-02	5	4	G[AG][GT][CG][CA][TG][CG][GT]T[GCT][CT]A[CG]TC[GA][AT][GCT][CGT][AC][AT]G[CT][CT]G[GA][TAG][GT][GC][GCT][TAG]A[ACGTA][TG]C[AG]
309	6.00E+00	2.10E-01	13	8	[ATC]T[ACGTA][GC][AC]TGA[AT][CA][AC][GA]C[GA][AG]G[GT][TC]
310	2.20E+00	6.80E-03	18	14	[GC][GA]TCC[GA]G[GC][AG][CA][GC]TC[GA][AGT]C[GC][AG][CG][GC][TA]
311	4.30E+00	7.00E-01	7	5	[CT]C[AGC][TG][GC][GC][ACT][TCG][GTC][ACT]CG[ACT][CG]G[TA]C[GCT][AGT]CG[CGA][ACG][AC][CT][GC][GA][AT][CG][CG][GTA]G[ACG][CG][CT][GAT][GT][CT][CA][ACG][GT][CAT][CG][GC][CGT][TC][TG][CG][TAG]
312	9.00E-01	5.30E-03	10	5	[ACT][GTA][GT][AT][GC][ATG][GAT][ACG][ACT][TA]T[TG][GT][ACT][TC][GAT][GAT][CA][AGT][TA][CAG][CAG][CGA][CG][GACTA][AC][GA]A[GA][AGT][GTA][ATC][ATC][GTA][TG][CT][GT]C[CTA][AGC][ACG][ACG][TC][TC][GC][TG]
313	6.40E+00	5.20E-02	10	4	[TA][AT]C[ACGTA][GCT]TC[AG][CG][GAT][AT][GT][AC]A[CA][TAG][GT][GT][CT][AT][CT][AG][TAG][CAG]AA[ACT][ACGTA][TAG][CG][AT][CGT][TC]A
314	7.30E+00	4.90E-02	13	7	G[AC][AC][GAC]GTC[AC]T[CA]G[CA]ACTC[AG]TC

Network	*E*-value$_t$	*E*_value	Cistrons in network	Cistrons with motif	Consensus sequence
317	1.60E+02	2.80E-03	15	14	[GT][TCA][CG][CG][AG][GC][CG][GTC][CG]G[GCT]CG[AGC]C[CGATA]C[CG][GC][TGA][CT]G[AT][GCA][CG][AT][GC][CG][TCG][CG][GC][TCA][CG][GAC][GA][GC][GC][GCA][GC][GC][GA][CG]G[GCA][CG][GACTA][ACG][CT]G
318	7.80E+01	4.90E+00	18	4	[AT][TA][AC][AT]G[GT][ACT][AC][GT][TG][CG][AG][TC][TG]T[GT][AC]A[CG][TG][ACG][AT][CG][CAG][TG][AC]
319	7.40E+01	1.20E+01	7	4	[AG][AG]G[AT][AG][AC]G[ACG][CGT][AC]A[TAC][AT][GA][AG]A[CG]G[GT][CAG][AG][AGT][CG][TCG][TA]C[TC]
320	5.50E+01	3.60E+01	7	7	[TA]TC[AC]CCGGT[CT][TC]CA
323	2.00E+00	5.00E-01	15	10	[TG]C[AGT][CT][GCT][TG][AT][CG][CT][TC]GG[AT][CTA][AG]
324	6.90E+00	5.40E+00	10	6	T[GT]C[CT]T[CG]GTCAT
325	1.40E+01	5.30E-03	21	3	GTCGT[TA][CG][TG]CC[AT][CG][TC]GCGTCGG[CA]G[TA]ACG[CG]G[TA][CA]G[AG][TG][CG]C[GT][CG][GA][GA]C[TC][TG]CTC[CG]TC
326	2.00E+00	1.20E-01	20	10	[TG]C[GA][GACTA][CG][GC][AC][CG][TGC][TA][CG]G[GC]C[GC][TC]G[TG][CAG][GCT][AGT][TA]C[GA][CG]CGA[CG]
329	2.70E+01	2.30E+01	8	5	[TC][GC][GT][ACT][CAG][GAC][ACT][GC][TA][GAT][ACG][GACTA]C[CG][TAG][ATC]G[GC][AG][GC][TA][AGC][CGA][TA][GACTA][GT][GA][ATG][GC][TC][GAC][CG][GA][GTA]G[GC][AC][GC][CG][AC]
330	6.50E-02	5.70E-02	8	6	[AG][AT]G[AC]AGGTCAT[GC]GA
332	7.30E-01	4.70E-03	32	23	[AC][GC][GAC]T[CG][GA][GTA]C[GC][CTG][GC][GA][CGT]C[CGT][TGA][CG][GA][CG][CG][GC][AC][CG][GCT][TA]C[GC][AG][GC][ACG][AT]C[CAG]T[CG]G[ACG]C[CG][AGT][GA][AC][TA][CG][CGA][AG][CG][GC]
333	1.50E+01	2.20E+00	11	11	[TA]C[GC][AC]CG[TC]CGT
334	3.10E+00	4.20E-04	7	7	GA[AC]C[AT]TT[CT]C[TA][GC][CT][ACT][TA]GT[TG][AT][CT]
335	4.90E+00	1.00E+00	7	5	A[AC][ACT][ATC][CA][TA][CG][AGT]C[TA]C[GC][AGT]T[GTC][AGT][GAC][ACT]T
339	3.90E+01	3.50E-01	19	6	[AG][CGT][GT]C[TC]T[CG][TA]T[CT]G[GT]TC[AC]C[CT]C[CA]GTG[CA][TA][TC]TCG[GC]TG[CT]TG[CG][TA]C[AGT]
340	2.10E+00	1.50E+00	8	8	[CT][TG]CGA[CG][CA][GA]C[GA]TCG[AC]
341	7.10E+00	4.60E-02	8	7	[AC]AGTTGC[AT]GA[TAC]
342	1.20E+00	1.10E-34	57	41	[GC][GC]TC[GA][AC]CG[AC][CGA][CG][AT][GC]G[TA]C[GC][ATG][CG][GC][TAC][CG][CG][AT][CG]G[AT][CG]G[ACT][CG][GC][GC][CG][GC][ATG]C[GC][TCA][CG]G
345	8.60E+00	6.10E-01	21	5	T[TC][GTA][TCG][CAG][GTC][TA][CTG][AC][GC][TC][TC][ACG][AGC][GA][TA][ATG][AGC]A[ACG][GC][ATG]TC[AT][AT][CA]
346	1.80E+00	1.50E-03	17	6	G[GT]TGAT[GC]GTG[AC][AG]GCAG[TG][CT]
351	1.20E+00	8.00E-01	15	11	[TC]G[CG][CT][GC][CGT][AT]C[AGC]CG[TG][TAC][CG][GC][GT]G[GC]C[GC][AT][TC][GC][CG]TC[ACG][CA]C[TA][TC]CGA[TC]CAG[GT][GTA]
354	6.50E+00	2.40E-01	25	16	CG[GT]CG[TAG]CC[GC]GGACG[ACG][CT]GA
355	4.40E+01	8.10E+00	10	9	GT[AG][CG]G[GT]GA[GC]T[GAT]
356	1.40E+01	8.80E-06	13	5	A[ACT][ACT][CT][TA][CG][AG][AC][CAG]A[TA][GC]A[AC][CA][AT][TAG][CA][AGC][CT][CTA][CT][TG][GCT][AGT][ATG][CGA][TA][TA]
357	8.90E+00	1.70E-02	10	8	AC[GC]A[CG][AG]TCG[AT]C
358	1.40E+01	4.50E-01	17	17	G[TG]CG[AC]C[CG][AG]G[GC][GAC]C[GC][AT]G
359	1.40E-01	2.60E-02	16	11	[CA][GT]AT[CG][TA]T[CG][AT][AGC][AGT]A[CT]C
360	1.60E+02	3.30E-01	7	6	CG[GA][AGT][CG]ATCACG[GA]TCA
364	9.00E-01	3.00E-05	12	5	AA[GT][TG]T[GT][GA]C[AT][AG][CT][CAG]A[AC]C[TC][CTG][ACGTC][AC][CG][TC][ATG][AT][CTA][CG][TAC][GTC][GACTA][TG]
365	5.00E+00	2.40E-09	21	8	GG[AG]CG[AC][CG][CG][AT]GC[AT][CG][GC]A[GA][CG][ACT]CG[AG][CT]G[GC][CG]C[AG][CGT]G[GA][CT][CG][GT]CG[GT][CA]G[ATC][CG]GA[CA]G[GC]CG[AG]
368	7.70E+00	9.40E-01	8	5	[TG][TG][CG]C[TG][CGT]CT[TG]C[GT][TC]G[CT][GC][TG][TG]

Network	*E*-value$_t$	*E*_value	Cistrons in network	Cistrons with motif	Consensus sequence
370	3.40E-01	1.30E-02	30	27	[TC][GC]G[CA]G[AT][TA]C[AG][TC][GCT]
373	2.40E+00	9.00E-02	7	4	[GT][AT][CGT][CAG][AC][TA][CT][GA][TA]CA[CG][CA][CAT][GA][TA][AG][AC][GCT][AC][TA][CG][AT][GT][CA][CA][GT][TG][GA][GT]T
374	1.60E+01	9.50E+00	16	4	A[AG][CT]AC[CAT][ACGTA][GAC]A[AC]C[TAG][CG][TA][CAT][GAC]ATCT
375	8.30E+01	2.60E-01	14	7	[TC][CAG]ATG[GA]CG[TA][GT]C[CGT]G[CA]T[CA][CA]TG[AC][GT]G[GC]T
379	1.20E+01	1.70E-01	4	4	T[AC]A[AC][CGT][TA][CA]T[AGT][CAT][CG]T[GT][AG][AT]
380	1.40E+00	1.20E-02	5	4	[AG][GC]G[AT][AC][AT][GT][CT][GAT][CAT][TG][ACT][CT][CT][CT][AG][TG][AG][TCG][TCG][CA][AT][CT][CA][AG][TA][GT][AT][AG][GCT][AT][GCT][TCG][GA][GT][CGT][CGT][ACT][GA][AT][TAG][AT][TG][CT][TCG][GT][AT]
383	2.80E+01	2.50E-02	6	6	T[CG]A[CT]C[ACT][CG][ACT][TA]C[CG][GC]G[AG]T[CG]AT
384	4.80E+00	1.10E+00	10	8	[CG]GG[AG][CAG][GT]A[CG][GC][GCT]C[GT][AT][GTC]G[TA][CA][CG][TC][GC][CG][AT]C[CGA]T[CAT][GC][ACT][ACGTA][CG][CG][CA]G[TCG][CG]C[GA][AC][GC][GA][CG][GA][CTG][ACGTA]G[CTA]G[CT][ATG][CT]
385	1.10E+00	4.90E-03	27	17	[TC][CT]G[CAT][CG][GC][AGC][CG]G[AT][CA][CG][TAG][CG][CG][AC][GA]G[TC][GC][GTA][TC]CG[AT]A[CG][GT][CA][CG][TC][CG][CAG][GCT][GTC][GT][GC]CG[TG]
388	8.50E+00	1.10E-03	16	16	G[AT][CGA][GC]TC[GC][TA]C[GT][GT]C[GC][TA][CG][CG][AGT][CTG]G[TG][CTA][CG][GC][CT][CG][AG][CA]G[TA]
389	4.70E+01	1.90E-05	32	16	[AT][CG]G[CTA]CGA[GC][GA][AT]C[CG][TG]C[GT][TA]CG[CA]C
390	1.60E+02	1.20E+00	23	17	GTGA[TC]CG[GA]
392	2.30E+01	1.50E-02	15	6	TG[CG][CT]C[GT]GGCTG[TA]CATGAC[GC][AT][CG]CAC[CG][TG][AT][AT]G
397	2.40E+01	2.30E+01	15	15	[GC]A[GA]G[TC]C[GA]T
398	1.20E+01	2.40E+00	4	2	TT[CG]T[AT]GA[AT]TAA[GT]T[CT][AT]C[AT]
399	1.60E+00	1.70E-01	8	4	[AT][GT][ACG][CG]G[AT][CA][AT][GT][TC][CG][AC][GT]G[AGT][AT][CT][AT][AC][ACT][AG][ACG][GT][GT][GT][CAG][TG][AC][AC][TA][TCG][ACGTA][GCT][TG][CAG][TA][CG][CAG]GG[AT]
400	1.40E+02	5.10E-06	24	13	G[AG][ACT]G[GC]C[CG]A[GCA][GA][CG][CG][GC][AG]GG[GAC][CT]G[ACG][CG][GC][AGT][CA][GC][GACTA][TC][GC][AG][TCG][GC][AGC]C[GC][GT]CG[TAC][TG]G[AT]
401	3.30E+01	1.70E+00	33	17	[TA][GC][GC][GAC]CG[AG]C[CG][GTA][CG]G[AT][GC][AC][CT][CG]G[CT][CG][GC][GT][GC][AGT][CT]C[GC][TA]CG[CT]C[CGT][AG][CT]GA[CG]
402	1.60E+02	1.40E-02	18	11	TG[CA]TG[CA][CT]G[GA][AT]CA[CG][CG]GC[GA][GC][ACG]GA
403	2.30E+01	7.40E-01	11	7	TC[AG]TCCT[GC][GC][TC][CG][AG]TC
406	3.10E-02	4.20E-03	29	5	[GT][TC][CT][GAT][GT][GCT][CG][AT][TA][GC][TG][CGT][GC][GC][CGT][AGT][CAGTA][AT][ATG][GTC][GAC][TCG][GC][CAT][ACG][TAG][CTA][AG][CAT]T[CT]A[GA][AT][GC][AG][AG]A[TA][ACG][AGT][AGC][CT][CTA][GC][GA][TACGA][TAC][ACT]G
407	2.50E-01	5.30E-02	8	8	A[GAC][GA][AG]C[TA][CAT]G[AT]T[GC][TC]C[GT][GA][TAG][CGA][AT]T[GT][AT]
412	7.70E-01	7.20E-01	25	11	[AC]G[TA][GAC]T[AGT]G[AG][GT][CG][AT][CT]G[TG][TG]TC[GC][GA]A[CA]
414	1.10E+01	2.60E-03	20	19	[AT]CGA[CG][GC]T[GC][CAG][CTA][GC]G[CGA]CG
415	9.60E+00	2.60E-03	22	11	[ATC]CG[GC][CG]G[AT]G[GTC]TCG[AGT][TC][CG][TA]CG[AT][TA]CA[CG][CAG][GTA][TC]G[AT]
417	5.80E-02	5.60E-02	10	5	[ATG][CG][TG][GAC][TA][GTA][ATC][TACGA][GT][GA][GC][CTG][GAC][AG][CAT][CGT][CAG][TA][CGT]G[GTC][TAG][GC][AT][GC][AG][GT][TA][GTA][CGT][GAC][GT][ACGTC][GAC][AGT][CT]GG[AT][TC][GCT][AC][CA][GT][GA]A[GTA][CGT][ACG][AG]
418	2.00E+01	3.40E+00	20	4	ATC[AG][AGT]T[AC][GA][CA][GA][AC][CG][AC][AG][CAT][ACG][GT][TCG][GCT][CAT]G[AT]A[AT][AT][AGT][CAT][GCT]A
419	1.40E+00	4.30E-01	7	7	CA[GC]G[CG][CG]G[GA]A[GAC][TG][CG]G[AG][CG]GAA[GC]

Network	*E*-value$_t$	*E*_value	Cistrons in network	Cistrons with motif	Consensus sequence
420	1.10E+01	9.90E-03	10	5	[AC][TAC][TC][CAT][GA][TA][ATG]C[AG][AG][GAC][TA][GA]CG[ACT][AG][TACGA][GA][TAG]T[TACGA][CTG][TC][CT][TCG][GAT][GCT][AC][GTA][TAG][CA][CTG][AGT][CG][GCT][GT][GT][GA][CT][CAG][TCG]T[ACT][GC][GA][TAC]
422	1.90E+00	8.80E-01	22	21	[GC][CAG][GC][CG][GT][CTG][GC][GT]CG[AT]C[GT][GT][TC][CG][TG]C[CG][AT][GCT][GC][CA][GC]G[GA][CT]C[AG]
427	1.50E+02	7.00E-02	8	7	[TG][CA][CG][CAG]GCAGC[AG][TG]CCG[CG]AG
428	2.50E+01	1.00E+01	10	4	[CT][CA][GA][TG]C[CGT][AC][AT]GA[TCG][TAC][ACGTA][GT][AGT][AG][CG][AG]A[ACGTA][TA][GCT][CAG][GT][ACT]G[TA]A[GAC][CGT][AT]G[TAC][CG][GC][ACT][GT][AG][CA]A
431	1.20E-01	5.30E-03	20	20	T[GT][TA]TCG[AC]
433	1.40E+01	1.70E-02	22	11	C[TCG][TG][TC]GA[CG]CT[GT][CG]AC[GA]T
434	9.10E+00	1.50E+00	8	5	[TC][TC][CG][GC][GC][AGT][AT][CA][AGT][CAT][GCT][AC][GTC][AGT][CA][ACG][CTA][TC][CG][AT][ACT][GCT][CA][GTA]G[AG]A[CAT][GTA][CGT][CAT][GA][TG][TC][ACGTC][CAT][GCT][TAC][TC][CAGTA][AT]
440	5.30E+00	2.30E-01	8	8	GTG[AT][AT][CT]C[CG]GT
442	2.00E+01	8.50E-01	28	7	[TG][AG][CG][GA][GT][CT][CT][TG][CA]GT[GA]A[AC]C[GT][AG][CT][GC][GT][TG][TA][ACG][ACG]G[TA][TA]GC
444	8.20E+00	9.00E-04	24	20	CG[GTA][GC][CG][AC][GA]G[TC]C[GCA][ACTGG][GC][GC]ACG[ACG][GC][GC][AT][CG]C[TA][GC][GC][TGC][CG][CG]
445	1.40E-01	1.10E-01	7	3	G[AT]CCA[CA][CT]G[TA]TC[ACT]A[TC][TA][AGT]T[GA][GA][AC][AG]C[GC]GTG[GA]T[CT][GC][TC][GT][ACT][CG][TA][GA][GC]A[ACT][AC][ACG][GA][GT][TA]G
446	5.40E+01	3.30E-05	10	6	C[GT]TG[AC]C[CG]TC[AG]AGC[AC]GG[CGT][TC][CT]G[AG]GGT
450	9.40E-01	3.20E-01	8	5	[AG][AC][AGT][TA][TC]C[TC][CA][GA][TC][AG][ACT][AT][ACT][TC][TC]G[CAT][TA][CGT][AG]
451	5.60E+00	8.50E-03	21	5	[AT][AG][CAGTA][TC]G[TACGA][TA][CAGTA][ACG][ATG][GAC][TA][TA][TA][TC][GACTA][TC][AT][AGT][TA][GACTA]AC[TAC][GCT][CG][AT]
452	3.70E+00	9.90E-01	13	13	[CG][TA][GC][CA][TG][GC][GTC]TG[CAT][TG][GC][GA]CC
453	4.50E-01	1.00E-01	10	5	[AT][CGA][GTA][TC][CTA][GACTA][ACG][AC][CT]T[CTA]G[AC]T[GC]T[ACGTC][AC][CGT]G[CTG][ACT][GA][AG][TA][TG]
455	5.40E-01	5.20E-03	24	17	A[CA][CG][CT][TC][GT][GC][TA][GAT]A[GTC][CG]AT[GA]A[CGT]C[AG][CG][TC]
456	2.20E+02	4.10E+01	12	5	[TC][CA]T[GC][GT][TA]T[TCG][ACT][CAT][CGT][GTC]TC[AG][TC][GC][AC][AC]
457	2.50E+01	7.10E-01	13	11	TCG[TG]CAC[CG]CG[GA][AC]A
458	9.90E+00	2.00E-01	14	5	[CT]G[TG][TAC][AT][TC][CG][GT][AG][GTA][AT][AG][CA][ATG]T[CA][CTG][TACGA][CGT][ACT][GT][CG][AC][AT][TG]A
459	5.50E-01	7.80E-02	12	5	[TA]G[AC][ACG][AGT][ACT]A[AGC]T[TG][ACGTC][CT][CGT][GACTA][CT][TC][CTG][AC][CGT][GT][TAC][AT][CAGTA][GACTA][ACG]C[AGT][TC][CT][GC][CA][TC][CAT][TG][ACG]
461	3.90E+00	1.10E-03	22	16	[TAC][GC][GAC][CA]G[ACG][AT]C[AGT][AC][CG]G[TA]G[CT][CA][GC]ATC[AG][GTC][CG][GT][ATC][CG]GA
462	9.90E+01	2.50E-13	43	43	[CG][AT][GC][CG][CTG][CG][GC][ATG]C[GC][AT][CG][GC][TAC][GC][CG][TG][CG]G[GA]C[GC][GA][CG][GC][TAG][CG][CG][AT][CG][CG][GCA][CGT][GC][AG][CGA][GC][AGC][CG]G[GT]
464	9.50E+00	7.50E+00	16	7	[AG]T[GA]A[ACT]GAAAC[GA][TC]TGA
465	1.60E+00	2.20E-03	9	7	[AT][ACG][CG]A[GT]C[AC][AG]G[AG][TA]C[AT][CG][GT][GA]T[GAC][CG]TCAA[AC]C[AT][CG][ACG][TA]
466	1.10E+00	3.90E-01	9	4	A[AG][AG]C[AG]C[CGT]C[AG][AT][TA][CT][GT][CG][CT][AG][GAT][GAC][TG][CG][TG]T[GT][CGT][AC][CT][AT]T
469	3.90E-01	6.50E-02	12	12	[GT][AC][CGT]GTCGAT[GT][CA]
470	9.40E-01	1.90E-05	11	2	TACTGTCGAT[AG]TCAGTTGCAGTTGTGGTTCCCGAA[AT]TT[GT]CA
471	6.70E+01	1.60E-02	10	10	[GC][GAC][CA][GC][GA][CTA][GC][GAC][TC][GC][GAT][TGC][CG]G[CAG][GC][TGA][CTA]G[GA][TG]C[ACT]CG[AGC][TC]G[TCG]
472	2.50E-01	1.00E-02	41	6	[AC][CA][AT][AT]TC[AC]AC[ACG][AT][CT]TT[GA]TC[AT]AC[AG]CT[CG]CTT

Network	*E*-value$_t$	*E*_value	Cistrons in network	Cistrons with motif	Consensus sequence
473	1.20E+01	3.00E-03	26	6	GC[AGT][CT][AG][AG][AT][CG]ATTT[CT]C[GA][CT][AT][AGT][GC]TCA[AG][CA][AG][AC][GT]A[AGT]TG[TA]
475	2.40E-01	3.60E-04	8	5	[TC][CT][ACG][TACGA]G[TC][TAC][ACGTC][TC][CA][AG][CAG][TAC][ACT][TACGA][CTG][GT][AG]T[ACG][AT]C[GACTA][GAT][TC][CT][TA]C[TG][TG][AGC][ACGTC][AT][AT][GAC][AT][GC][TA]
476	8.30E+01	1.90E+01	13	7	TGG[TA]CG[CT]CTTC
477	4.80E+00	5.50E-01	20	8	[GA][CGT][CTG]G[ACT][CG]G[TCG]CG[GCT][CG][CG][GA][CG][CG][AT][TC]GA[TC][GC][AC][CT][GC][AGT][GC][GC][AG][TA][GC]A[CG]GAG[AGC]T[CTA]
478	2.60E-01	1.70E-02	11	9	[CA][TCG]G[GA]A[CTA][GT][AC][TG][GAC][AT][TC]T[CT]C
479	8.80E+02	1.80E-01	52	7	[AT][TC][GT][AC][TA][GAT][TA][TC][CG]A[AC][TG][AGT]TT[CG]A[AC][CA][CA]A
481	4.00E+00	3.30E-01	8	8	G[GT]G[AT]G[AT][CT]C[AG]T
482	3.00E-01	1.10E-01	11	5	[TA][CG][TAG][TACGA][CT][GA][ACG][CA][AGC]A[GC][AT][CG][GC][GCT][TG]G[AT][TC][CGT][AGT]CG[AT][TC][GACTA][CTG][GTC][CA]G[AG][GC][TC]
483	1.40E+01	1.70E-01	11	6	T[AC][TG]CC[AG]T[AG]G[ACG][TG][GC]A[AC][CG]A[GA][GT]AC[CG]CA[CG][TA]G[GT][AT]
485	2.20E+00	9.20E-02	13	11	G[CG]TC[CG]GC[GC][AT][CG]G[ACT][TG][CG][CA]GCG[GTC][GC]A[CA]GGCG[TAC][CG]G[GA]A[GCT][GA][GAC][CG][GA][CGA]C[AG][GA]
486	6.80E+00	2.50E+00	9	9	G[GA][CTG][GAC][CAG]GGA[CG][GC]AG[GTA]TC
488	2.10E+01	6.70E+00	11	3	T[GC][AT][TA][GA]GT[AT]TG[TA][TG]C[AT][TA]
490	1.10E+01	1.80E-01	14	8	[CG][AGT]T[CGT][ACG][AC][GC][GCT][TA][CG][GAC][AG]G[GAT][AC][CG][TG][AG]CG[AC][GC][ACGTA][CG][CG][CA][GTA][CG][CA][TC][GT][CAT][TG]C[GC][CGT]C[AGC][AG]G[CG][TA]C[GC][GT][CG][ACG][TAC]
491	2.00E+01	3.60E+00	11	7	[TA][AT]C[GA][TA][GC]ATC[TCG][TA][CG]TAC
492	1.90E+00	1.80E-01	13	13	[GC]A[GC][CG]AGG[TA]CG[AT]
493	1.00E+02	2.40E+01	8	7	GT[CG]G[AT][CT]C[TG]G[AT]T
494	2.20E+00	1.20E-03	21	21	[TCG]T[CG][CG][AG][CT][CG][AT][CG]G[AG][CG][CG][TCG][GCA]G[GC][CG][GC][AC][TCG][CGT][GTC]G[GC][AC][GC]G[TA]C[CG][TG][CG][GCT][TAG][CA][GC][GA][TC]G[AT]
497	4.60E+01	6.20E+00	10	5	[AG][CT][CG][AGT][CA][GC][AT][AGC][CGA][TA][GCT][CG][TC][GT]G[AT][GT][GACTA][TA][CGT][GA][CA][GT][GC][AGT][CGT][CAGTA][TACGA][GC][CG][AC][GAT]G[AT][CG][GT][TC][AC][CG][AT][ACG]
499	5.80E+02	2.30E-03	42	18	G[CT][CG][CG][TG]CG[CGT]C[GC]A[CGT][CG][GC][CG]C[TG]C[GC][AC]G[GC]A[ACGTG][CG][CAT]G[CG]A[CG][CG][GT][CT][CG][CG][GC][CG][GTC][TG][CG][TG][CG][GC]G[CA][GC][AG][CGT][GC]
500	2.20E+02	3.90E+00	12	12	GG[AT][CG][CG]G[GC]G[GA][GC]GAG[AG]A
501	2.60E+00	7.10E-01	10	5	TC[TG][TC][GAT]G[AT][CGT][CGT]TGA[AT]
502	7.20E-02	3.20E-03	17	5	C[GTA][TG][TACGA]A[CAG][TG][CTA][CG][AGC][AGC][ACG][CTA][GCT][TG]T[TC][CGT]T[GA][AC][ATG][ACT][CAGTA][TC][TA][CAGTA][TAC][TAG][GTA][ACG][CT][TACGA][TA]T[CA]C[GT][CAGTA][AT]
505	2.60E+01	1.60E+00	12	8	[CA][TA]CG[AC][CG]C[AG][CG][CG][AT][CG][CT][AT][GC]G[ACG][CA][GC][GTA]C[CA][CGT][CG]G[AC]C[CG]A[CT][CG][TC]CGA[GC][CT][GA]G
509	2.20E+00	5.50E-01	6	6	CCTT[CGT]TGC[GA][ACG][AC][CG][AG][AT]T[GT][TC]
512	1.50E+00	4.00E-01	13	9	[TA]CG[AGC]C[CAGTA][CT]G[GT][AG][CGT][CA]TG[CGA][TAC][GC][CAG][TA][GC][AT][TCG][CG][CG][TA]C[CG]T[GC][ATC][CG][CG]G[GAC]C[GA]
514	8.80E+01	2.90E+01	13	4	[AG]C[AC][ACG][GT][GCT][TA][GAC][AG][CG][CAT][AC][AG]G[AT]TC[CAG][ACGTA][CG]AC[CA][AT][AG][AG][ACT][CG][TAG][GA][TG][GAT][CT][ACGTA][GA][GCT][AT][AC]C[CT]G

Network	*E*-value$_t$	*E*_value	Cistrons in network	Cistrons with motif	Consensus sequence
515	2.30E+00	6.50E-02	12	5	[TC][CG]GTC[CG][AGT][ACT]G[TA][AT][GA][ATC][AC][GA][CGT][GCT][GC]A[GC][CTG]G[AT][CG][CAT][CGT][GC][AT][CGA][GC][CGT][TG][CGA][AT][CT]
516	6.60E+01	4.50E-02	11	9	[TA][CG][GC][TA]C[GCT][TCA]C[GC]TC[GT][CGT][CT][GC][ACG][CG][CA][GT]G[GC][ACT][CG][CGT][ATC]C[CA][ATG][CG][GC][ACG][CG][CT][TC][GC][GAT][GACTA]C[CTG]
517	3.20E+01	1.30E+01	14	10	C[GA]TC[AGC][CG][CG][GA][AG]C[ATC][TC][CG][AGT]TC[CG][GTA]C[TCA]G
518	1.20E+01	4.60E+00	17	4	T[CAT][GAT]C[TA]CA[TA][GCT][ACGTA][AG]A[CAG][ACG][CAT][AG][TA][GAT]T[ACGTA][CAG]G[CT][TG][TC][CA][AG]G[TG]
519	1.40E+01	3.90E-02	5	4	T[TG][CG][AT][GCT]A[AT][CG][GT][CAT][GC][TG][AT][CAG]A[AG][GC][AG][TA][CGT][CGT][ACG][CG][TAG][TG]C[GT][ACGTA][AG][AC][CAG]G[GT][CG][GT][GCT][AG][CG]G[CGT]G[GT][AG]
520	1.00E+01	4.10E+00	13	8	T[CG]TTCA
521	5.80E+00	2.00E-01	9	6	T[TC]G[GC]T[CG][TA][AT][GA]T[CT][CT][AC]C[AT][GT]TG[AT]G[GT]C[CG]GA[AC][CG]T[GT]
523	9.70E+00	4.00E-01	14	8	A[CA][TC]C[CG][GT]TCGAG[GC][TA]G[GA]T[TC]
525	2.90E+00	1.20E+00	5	5	[TC][GC][TAG]TC[ACG]CCA[TG][TC]
526	1.50E+02	2.90E+00	14	10	AGG[CG][CT][GCT][TC]C[GA][ACG][CT]GA[CA]G
530	3.60E+00	2.20E+00	12	4	[TC][TCG][CAG][AG][ACT]CG[GT]G[CT][AG]A[GT]G[AC][ACGTA][TG]T[CGT][CGT][AC][CG]G[TA][CG][GA][ACT][GC][TG][CAT]G[AT]TCG[AG]
531	1.80E+01	6.60E-05	35	35	[CG][GTC][TAC][CG][GC][AC][GC][GCA][ACGTG][CG][CG][GAC][GC][ACG][CGT][GC][CAT]TCG[GCA][CG][GT][CAT][GC]G[TC][CG]G[CA][GC][GA][TCG][CG][GA][CTA][CG][GC]
532	4.70E+00	3.30E-02	29	8	A[AT]C[GTA][TA][CG][CGA]AG[GAT][TG][CG][GC]CCG[AT]C[AG][CA][GC]G[TC][CG][TAC][CGT][CG][CT][AT]CC[TA]C[CA][TC][CT][CT]TC
534	4.20E+00	2.40E-01	15	14	[GT][CT][CG][AC][GT][GA][AT][GTC][CG][TA][CG][CG][TA][CT]GA[AT]
535	6.10E+00	1.60E+00	6	5	[TA]T[GA][TC][GC][AC][AT][CTG][ACGTC][TC][GT][GT][CT]A[ACT]G[GA][GA][GTA][ACT]
537	6.00E-01	2.90E-03	17	15	[CG][CG][TA][CGT]G[TG][CT][GC]A[GC][GC][AG][CT]G[AGT][AC][CG][TAG][GC][CG][GC][CAG][GC][AT][CA][GC][GCT][CT][CG][AT][GC]C[ACT][CG][CG][TCG][GTC][GC][AT][CGT][CGT][GTACA][CAG]CG[GT][CG][GAT][CT][CG]
538	2.50E-01	8.20E-02	9	7	CG[GAC][CT][CA][GCT][CT][AC][TG]TC[AT]C[CAG]ACGA[AC][GT][AC][CG]C[TA][CAT][TC][TC][GC]CG[CT]GCC[AG][CT][GT][CG][TC]GG[TC][GCT][GAT][AC]
539	9.10E+00	6.10E+00	9	5	C[TA]C[GC][GACTA]CC[AT]G[GC][AG]G[CTG][AC][CG][GC][CAGTA]C[GC][ACG][GACTA][CG][AT][GTA][CGA][CAT][CG]C[TA][CG][GC]A[AGC][TC]
540	4.70E+01	4.60E-04	19	12	[AT]C[CG][CT]G[GC][CT][GC][CG][TG][GC][CAG][CG][CG]G[AGT][GC][GCT][TC][CG][TG][TC][CG][ATG][TC]C[CA][CA][CG][GC]TCG[CG][CG]G[AG]C[CG][CT]G
541	4.70E-01	6.00E-02	17	17	G[CTA]CGA[CA][CA][TC]C[GA][CT][CG][CG]A[GC]
542	9.70E-01	2.30E-03	8	5	T[ACG][AGT][GAC][ACT][TA][AC][AC][ACG][ACG][AGT][CGT][TC][GA][ACGTC][TAG][CT][AGC][GT][TAG][CGT]T[CGT][TA][GTA][TC][CT][TACGA][TG][AGC][AT]C[ACG][TA][TA]
544	1.40E-03	6.60E-06	9	9	A[TA][GA][CA][AC][GTA][AG][CAT][TC]GA[CAGTA]C[AG][GA][CGT][CT][CT][GC]T[TC]
549	2.40E-02	9.80E-05	7	4	[GT]A[ACT][AG][TC][GCT][AG][AGT][CT]T[CA]T[AG][CGT][ACGTA][AG]AT[GT][AG][TAC][ACT][AG][CAT][CA]A[TA][GCT][TC][AC][TC][AT][CT]
552	1.90E+01	1.10E+00	6	5	G[TG][TA][GC][GTC][GACTA][GAC][GT][CTA]G[AT][TA]G[AT][TA][CG][AG][GC]C
553	3.00E+01	3.60E-01	14	2	A[CT][CT]G[AT]CTACA[AG]T[CT]C[AT]GTAGAC[CG]GTCTAC[AT]G[AG]A[CT]TGTAG[AT]C[AG][AG]T
555	6.70E-01	2.60E-01	11	8	GT[CG][GC][TAC][CA][CG][AG]G[GT][TA]C[GC][CT]C[ACGTA][AGT]T[GT][TC][CG]G[AG][GC][TG][CG][ACG][GCT]T

Network	*E*-value$_t$	*E*_value	Cistrons in network	Cistrons with motif	Consensus sequence
561	4.50E+00	3.70E+00	6	4	[CT][AC]C[AC][AT][CT][CT][GC][ACGTA]A[CG][TA][CG][TAG][GT][CG][CT][CG][AG][ACT][AT][CA][GA]C[CG][CGT][TA][CAT][GCT][AG][CAG][CT][CG][AC][AC][AC][GAC][GT][TA][CT][GA][GT][TC][GAT][AGT][AG][CT][GCT]G[AT]
562	5.90E+00	1.70E+00	6	5	C[ACG][TG][GT][AG][TC][TG]G[TC][CA]GT[AC][AT]
563	3.30E+01	1.50E+01	6	4	[GT][TC][CG]A[GT][CG][AT][CAG][CT][TA][CG][AG][GT][CG][CT][GT][TC][GC][GT][CT][GC]A[GT][GT]A[GC][GT][GAC][TA]
566	2.60E+01	1.60E-02	19	13	C[GC][GT][AC]CG[TC]C[GC][CTA]G[GC]A[AGC][CG][TG][GC]C[TG][CG]G
567	1.10E+01	6.60E-03	22	22	C[GCT][TAC][CG][GC][ACG][GC][CA][CA][CG][GC]T[GC][CG][TG][CG][GC][AGT][GAC][GC][AC][GC]G[TA]CG[GC][CG]G
569	4.30E+01	7.70E-03	19	12	TC[AC][TC][CG]A[CT]G[GA][AT]CA[CA][CG]
572	1.00E+01	6.40E-01	6	5	[TA][TA][CGT][TCG]C[ATG][CG][GA][AG][CT][TA][GC][CGA][TA][TC][CAG][TA]C[GC][AGC][CT][GA]C
575	6.10E+00	3.20E-02	9	4	[TC][TA][CA][CG][GAT][TA][GA][TC]GA[AT][CT][TA][AC][TAC][TAC]GA[AT][AT]
576	4.90E+00	2.80E-03	25	12	[TAG][GC][AC][AG]G[GA][AC][GC]C[CT][GT][TCG]CGA[GAT][GC][CG]CG[TG][CT]G[AC][TA]G[TA][CT][CG][GCT][CG][GC]A[ATC]G[AC][ACT][GC][AC][CA][GC][AG][GC]G[TC][GC][GC][GT]C
579	4.00E+00	8.40E-02	21	21	[AGT]CG[AG][CG]G[GTA]C[CG][GA][GC]C[AGC]CG[GC][CG]G[AT][CA][CG][GA][CG][GC][GAC][CT][CG][GA][CT][GC][AT][CG][CG][GA][CGT][GC][TA][CT]
581	4.30E+00	2.80E+00	8	4	GC[GAC][GT]T[CG]A[GT][CG][GT][AT][AT][CT][GC][TG]T[CT][AT][TA]
582	7.60E+00	2.80E-02	27	5	[ACG][TC][TCG][AGC]C[GC][TA][TG][TCG]T[TC][GTC][TAG]T[GACTA][CG][AGC][CT][TAC][CG][GT][TAG][TC][TG][TG][GC][TCG][ACG][GT]GT[AGT][CT][TACGA][CT]G[GT][GAC][CAGTA][AG]
585	1.20E+01	1.10E+01	12	11	[AGT]GGT[CA][GA][AG][CG][CAG]CC[CG]TCG
589	3.10E+01	6.70E-03	29	6	[AT][CG][CT]T[CG]GT[CT]G[AC]G[CG][AT][GA][CT][CGT][GC]C[CG]TG[TA][CT][CG][ACG]G[GT][CG]C[TG]T[CG][GA]A[AGT][GC]T[CGT][GA]TC[GC][AG]T[GT]C[GT]GT
590	4.10E+01	3.00E-02	10	8	[TA][CG][GAC][AT][GT][CG]C[GC]G[AT][GC][GA]C[GC][GT]C[CG][ACG][CTA][CG]G[CTA][CG][AT][TAC][CG][GT][ATC][CA][GC]CG[GT][ACT][CG][CGA][AT][CA][CG][TA][GC]G[TA][CG][AT][GTA]C[GA]TC
591	2.20E+01	1.30E-01	9	9	G[TA]G[AT][TA][CG]GAC[TG]
592	3.00E+00	3.60E-02	20	4	G[TC][CT][CAT][CG]A[GA][AG][TG][CAT][TAG][CAT][GC][GT][CGT][TA][CT][CAT][AG][CT][GAC][AT][TA][ACG]CGG[AT][AG][AG][CT][AT][CA][GC][CT][TC]T[CG][TCG][ACGTA][CA][TAG][AT][ACG][AG][AT]
593	2.30E-01	3.40E-02	7	7	GGC[CT]TCG[AC][ACG]G[AGT]C[CG][AG][CT]C[CG]GGAT
597	2.20E+00	4.30E-02	15	5	GG[TAC][GC][AT][GTA]C[GC][TC][CT][ACT][TA][CT][GA][AT][GC][TCG][CAGTA]CA[CGA][TC][GAC][CTA][CAG][CAT][ACT][GC][GT][TCG][CG][ACT][TA]G[TCG][CAT][CGT][AC][AT]
599	5.50E+01	7.50E-01	16	5	[CA][TC][GCT][CAT][TC][GTC][GCT][CGT][CT][GA][AG][TAC][TACGA][GT][GCT][GCT][GT][CA][GAC]A[CT][TCG][CGT][GT][GACTA]TG[CT]T[TCG][CA][CT][ACT][CT]TC[CGT][TG][CGT][TC][TACGA][CGT][ATG][GCT][AGT][AC][AT]
600	4.60E+01	3.80E-01	21	6	[TA]C[AT]T[CG][TC][CG][GC]TG[AT][TG]AA[AG][AC][GT]C[AG][CA]CGCAC[AT][TA]C
602	1.30E+00	2.90E-04	22	22	A[CAG][GT][TC][CT][GC]TC[GCT][AT]C[CG][GA][TCG]GA[CG][CGA][TGC][CTG][GC][TAC][GC]
606	2.80E+00	4.40E-01	11	6	T[GC][CT][AT][CG]GACTTC[GA][AC]
608	4.50E+00	1.00E+00	14	14	[TC]CGTC[GC][GT][CTG][GC][AC]
609	4.50E+00	1.30E-03	11	10	[CAT][CT]GA[GTA][CG][TCA][CG][GCA][TAC][GC][GT][TG][CG][GAC][AT][GC][CGT][AGT][CG][GAC][TG][GC]G[AGCTC][TC][CG][TA][GT][GC][AG][GCT][GC][CAT][CG][GCT]T[GC][GCA][TA]
610	2.40E+02	4.10E+00	17	8	C[AG][CA]GG[TC][GC][AC]TG[AT][CT][GC][AG][CA]
611	1.40E+01	4.40E+00	9	7	G[TA]T[CA]G[CT][CG][AC][AC][GA]T[TCG]C[CG]CGT[AT][CT]G[TC]

Network	*E*-value$_t$	*E*_value	Cistrons in network	Cistrons with motif	Consensus sequence
613	6.30E+01	4.60E-02	12	8	C[GT][ATC]C[CG][CT]CG[AC][CA]GAC[GC][AG]G[CGT][TG][CG][GC][GTC][CG][CG][TCG]C[GAC][AG]C[GC][GT][CG]G[TG]C[GT]ACG
618	4.10E-01	2.00E-01	6	5	CC[AC][TC][AGC][ATG][CG][AG][CAG][GTC][GT]A[TA][GACTA][AT]T[CT][GTC][ACG][TA]
620	3.20E+00	9.40E-01	14	7	[AT][GC]GAC[CT][TC]C[CAG][AT]GC[AGT]C[GA][GCT]C[CG]T[GC][CG]G[CG]GAC[CAG][TG][GT][GA][AC][GC][CG]G[GC][GC][CA][CT]C[ACT][CG][GA][AG][CG]GA
621	6.30E+00	3.20E+00	26	10	GA[GC]CTGTTC[CG][TG]
622	7.50E-01	1.30E-01	21	5	[TG]C[TG][GA][TA][CA][CGT][CGA][ACG][AGT][CT][CTA][CG]C[GA][TA][TG][TG][CT][CT][GC][CT][AG][CGT]T[GC][GTC]A[TC][GT][AC]C[GTC][TG][CT][TACGA][CAGTA][CGT][CA][TC]
623	3.50E-01	7.50E-02	10	5	[AG][CAT][GTA][ATG][TAG][CA][TA]C[TACGA][ACGTC][CGT][GC][GAT][ATC][ACT][CAG][TC][CGT][ACG][GACTA][CTA][TA][TG][TAC][TG][TG][GTC][GACTA][ACT][CAG][AT][GC][TC][TC][CTG][AT][ACG][AC][TAG][GAC][AT][AT]
626	3.20E+01	6.90E-01	17	11	[AT][AG][GC][CAG]GG[ACG][GCT][GC][TC]C[GC][GAT][CG][CG][AT][CT]G[CT][CT][CG][TAC]CG[GT]CG[TC][CG][CG]CG[CG][AGT]C[GC][TA]C[CG]T[GC][GT]TC[GC]T[CG]
627	3.10E-01	7.00E-02	10	10	C[CG]T[CG]G[AG][CG]G[GT][GT][AG][TG][CT][GA][GA][CGT]C[TG][GT][CA]T
628	1.50E+01	6.10E-03	27	22	[AT][CG]G[AG]C[CG][TA][GC][CA][TC][CG]G[CT][CG]G[GA]CC[CAG][GC][GC][CTG][CG][GA][TG]C[CT][TAG][CG]G[ATC]C[CG][AT][CG][CG][AGT][CG][GC][ACT][CG][GA][TC][CG]G[CT][CG]G
630	1.10E+01	2.20E+00	9	6	A[CG][ACG]AGG[AG]AGG[AG]G[TA][GT]GAC
633	7.60E+00	2.70E+00	10	10	[AT][TC]GA[GAC][GTC][AT]CG[GT][CTA]
634	1.40E+02	8.90E+00	14	14	[CG][CT][TA][CG]G[AGCTC]C[GC][ACG][CG]C[TCA]GG[CA][GC]C[ACT]GCA
635	6.10E+00	1.00E+00	14	8	[TC][CT]C[AT][CTA]G[TG][CA][CA]T[CGT]C[TA]C[GC][GT]T[AT]C[ACG][AG]
636	3.40E+01	1.40E+00	9	9	[TC][GC][AG]T[GC][TA][AT]CA
637	9.50E+00	3.40E+00	15	15	[TG]CCGG[AGT][CAT]G[GA]CG[CT]
638	8.40E+00	1.00E-07	20	18	[TA]G[GA]T[GC][GAT][CT][CG]G[ACG][CG][GA]TG
639	1.60E+02	3.10E-01	19	15	[TG][CA][CG][ATG][CA]G[ACT][CA]G[ACG][TGC][CG][TAG][CT][CG][TG][CA][GC][GTC][TCG][CG][GTC]C[GC][GA][AT][GC][GA]
641	1.20E+00	1.10E-02	13	7	C[GA]AC[AGT][CGT]C[CA]CG[CGT]TC[GT][TA][CG][AC][CGT][CG][CAG]T[CG][AGT][AC][GC][AT][AC][CG][GC]T[GT]C[TG]C[AG][CT][GA][GC][TC][GC]CT[CG][GC][TA]C[AG]
645	2.70E+00	6.10E-06	8	4	[TA][CGT][CT][TAC][GAT][AT][TG][CGT][CT][TC][GT][TAG][CT][GA][ACT][GCT][ACGTA][CA][GA][AT][CG]A[AT][TAC][AC][TG][ACGTA][CT][GC][CA][CAG][AG][ACT]T[TA][TA][TC][GC]T[CT][TG][AT][GAC][GT][CG][AT][ACT][CA]A[AG]
646	1.30E+02	1.90E-01	26	26	C[CG][TG]TC[CAT][TGA]C[GA][TG]C
647	4.20E+01	1.40E-01	17	17	[TC]CGTC[CA]T[GCT][GA][CT]C
648	1.30E+02	1.00E+00	8	8	T[GA][GT][AC]CGTCAC[CG][GA][CA][CT]GT[CG][AG]
651	2.80E+01	6.60E+00	9	6	TCAGT[AG]C[AG][AT][CG]CGCGT
652	1.10E+02	1.10E-01	12	12	G[AT][CA][GC]AC[GC][TC][GC]G[GC]G[AGC][CA]CGAC[CG][AC][CG]
653	9.80E+01	6.80E-03	10	10	C[AC][TC][GC][TAC][GT]G[ACT][CGT]C[AC][GC][GT][AT]C[GC][GCT][TAC][CT]A[GA]
654	6.10E-01	1.70E-01	13	6	G[TC][CG][TC]TG[AG]T[CG][AC][CG][GC]A[AG]TG[TA]T[CG]A[TG]C[AC][AG][GT]TC[CG]A
655	9.70E+00	2.20E+00	7	6	[AC]ACT[GC]GA[AC]GGC[ACG]T[GT][CG][AC][CT]G[AT]C
656	4.60E+00	1.30E-03	12	5	[TCG][ATG][AC][TA][CT][CTG][ACG][CGT][TG][CG][CAT][ATC]C[AT][ACG][CTG][TA][AT][ATG][TACGA]C[AG][ACG][AT][GAC][ATC][GA][GTC][TA][GAC][AC][GT][CTG][GA]T[TCG][AT]
657	1.00E+01	4.30E-01	12	5	T[CG][TACGA]TC[CT]TC[GCT][TA][GAC][GC][AGT][CAT]G[AG][CT]C[GTC][GAC][GCT][TC]CG[AGT][CA][GAC][TA]

Network	*E*-value$_t$	*E*_value	Cistrons in network	Cistrons with motif	Consensus sequence
660	8.80E-01	5.60E-01	20	6	CGA[AG][CT][ACG][CT][AG]A[CA][CT]C[AT][GA]AT[AT]A[CT]GGA[GT]TAT[CT]
661	5.90E-01	2.10E-01	9	3	[AT]T[AG][GT][AGT][ACT][GT][AT][AGT][AT]C[TA]C[AT]A[CA][TC][AT][TA][GC][AT]T[CGT][TG][TA][AGT][ACG][TA][ACT][AGT]C[TA][CA][CA][ACG]A[TG][TG][CG]AG
662	1.80E+00	2.60E-01	10	4	[AG][CT][GA]A[TCG][ACT][ACGTA]C[AT]G[AT][CA][AT][GCT][AG][CAG][CAT][GCT]G[TAG][GT][CAT][AT][TAG][TG]C[TA][GC][AG][ACT][ACGTA]T[CT][ACGTA][GT][CT]A[TG][GC][AGT][AT]
663	2.90E-01	6.60E-02	7	5	[AG][TAC][CGT][AGT][CGT][TACGA][GT][AG][AT][ACT]G[AT][CAG][CTG][CGA][CGT][TA][TCG][CT][AT][GC][ACGTC][CG][TG][GAC][AGC][AT][CAG][CTG][ACGTC][GAC][TA][CT][TAG][TC][CT][ACT][GTA][GA][TAC][CGA][ACGTC][CTA][TG][CAGTA][ATG]CA
665	2.80E+00	1.00E-01	12	9	[AG]T[TG]T[CT]CG[GC]GCT
666	1.50E+00	4.00E-01	9	6	[TA][CT]GTACG[GT]GA[TA]GT[AC]GA[TG]G[GT][CT]G
667	6.10E+01	1.40E+01	9	3	TG[AT][CGT]AG[ACT][AG][GC][ACG][GA][TG]C[TA][CA][AC][ACT][AGT][TA][ACG][AC]G[TA][TA][CG][ACT][TA]T[TA][TA][ACG][TG][CT][TC][AC][CT]A[AG][TC]
668	9.30E+01	6.40E-01	13	10	G[ACGTG]CG[AT][GCT][CG][ACGTG]CG[CG][CG]G[AC][CA][GA][ACT][CG][GC]A[GT][GA][TA][CGA]G[CGT][GAT]C[AG]
669	1.50E+00	8.50E-01	11	5	[GA][AG][TC][CA][GTA][TA][CT][TA][TC][CGT][GAT][TA][CT][GT]A[CG][CG]A[CGT]G[TA]
671	3.60E+01	5.70E-03	14	14	[CG][TG][CG]G[TAC][CT]G[ACG]GG[TAC][CG]G[GC][CT]G[AT][CAG][CGT][TAC][CG][CG][ACG][CG][CG][CA][CG][GCT][GC]C[GA][ACC][CA][AG][GC][TG][AGT]CG[TG][CG]C[TCA][CG][GT][TC][TCG]C
672	1.10E-01	7.70E-02	12	6	A[TC]GGA[AC]T[CG]A[AG]A[GA]T[GT][AG]A[CT][GT][CGT][TG][AC]T[AG][GA][TC][GT][GC][GT][AGT][AT][CG][CT]A[TA][CG][AC]
675	8.30E-01	6.10E-04	21	21	[TCA][CG][CG][GA]CG[TC]C[CG][TA][CG][GC][AC]C[GC][TG]GG[AT]CG
677	3.20E+01	2.30E-01	11	9	[AT][CAG][AG]TG[AT]T
678	4.80E+00	3.20E+00	11	11	G[AT][TC]G[CT]CGC[GT][GC]T[GC]C[TC][CG][CT][CAG]C[CG][TC][CG]G[GT][CT][GA][AT]C
679	3.40E-01	2.20E-02	9	8	TCG[TC]CGTCGA
680	1.10E+02	4.00E-02	11	7	C[AG]C[TAG][AGT][CG][CG][TAG][GC]G[TG]CG[AT][CG]C[AG]CGGC[AGT]TC[GA][AT]C
684	2.70E+00	1.10E-01	14	14	CGA[CG][CT][GTC]G[CT]AC[GT][AT][CAG][CG][TG]
685	2.00E+01	1.10E-02	10	10	GA[GC][AGC][ACT][CGA]GTC[AGCTC][TA][CT][ACT]T[CT]
687	2.90E-01	4.00E-04	10	4	[AT][AG][ACT]G[AT][AT][CA][TAG][AC][AT][GAT]G[AT][AG]A[GAT][GC][AC][GT][AT]A[CT][AC][TA][AT][TA]
689	1.90E+00	3.30E-02	22	5	[GT][CA][TAC][TA][ACT][CAG][GA][CAGTA][GCT][GA][TA][GCT][AC][TA][CGT][GT][TA][TAC][ACT][CGA]A[TAC][CGA][GTA][GC][ACT][GT][AG][AGT][ATG][CAT][ATC][TG][TG][TA][GC][ATC]A[TG]
691	1.30E+00	5.50E-01	9	9	[CA]CG[TC][TC]G[ACG]T[CG][AT]C
692	1.20E+01	1.50E+00	5	4	[AG][AG]T[CT][AT][CA][GT][AG][AT][CG]AT[GT][AT]

Appendix B

Mutations in a clavulanic acid high producer *Streptomyces clavuligerus* strain

Highlighted in blue are transporters, protein binding proteins, ribosomal proteins, and other regulators. Highlighted in green are genes which are part of secondary metabolite clusters.

B.1 Non silent mutations in coding sequences.

Table B.1. Non silent mutations in genes located in the plasmid.

Position	Reads	Locus	Annotation
18219	14	SCLAV_p0017	2-methylcitrate dehydratase
18220	14	SCLAV_p0017	2-methylcitrate dehydratase
47583	76	SCLAV_p0047	Hypothetical protein
433485	31	SCLAV_p0447	Hypothetical protein
561962	5	SCLAV_p0557	WD-40 repeat-containing protein
615002	4	SCLAV_p0601	Hypothetical protein
636423	4	SCLAV_p0621	urate catabolism protein
752810	7	SCLAV_p0711	Condensation domain protein
771761	11	SCLAV_p0727	Putative alcohol dehydrogenase
771762	11	SCLAV_p0727	Putative alcohol dehydrogenase
785221	26	SCLAV_p0743	Hypothetical protein
922868	348	SCLAV_p0866	Methicillin resistance protein
923196	408	SCLAV_p0866	Methicillin resistance protein
932783	406	SCLAV_p0876	Hypothetical protein
944324	389	SCLAV_p0887b	Putative integral membrane protein
1022876	323	SCLAV_p0959	regulatory protein
1234853	486	SCLAV_p1129	HPC2 multi-domain protein
1270488	430	SCLAV_p1149	Cupin 4 family protein
1270489	432	SCLAV_p1149	Cupin 4 family protein
1297373	371	SCLAV_p1172	Undecaprenyl pyrophosphate synthetase
1328094	466	SCLAV_p1198	Tail sheath protein
1440512	16	SCLAV_p1283	Moenomycin biosynthesis protein MoeGT4
1544237	8	SCLAV_p1355	Hypothetical protein
1558250	31	SCLAV_p1367	Peptidase
1572072	5	SCLAV_p1376	Subtilisin-like protease
1576092	3	SCLAV_p1379	plastocyanin
1576793	3	SCLAV_p1379	plastocyanin

Table B.2. Non silent mutations in chromosomal genes.

Position	Reads	Locus	Annotation
27770	806	SCLAV_0012	Modular polyketide synthase
63639	532	SCLAV_0015	Modular polyketide synthase
295811	727	SCLAV_0202	Putative non-hemolytic phospholipase C
519924	345	SCLAV_0380	Cell surface mucin-like protein
539751	687	SCLAV_0395	ABC-type multidrug transport system, ATPase and permease component
547483	868	SCLAV_0401	Putative siderophore-interacting protein
583151	718	SCLAV_0431	Hypothetical protein
820316	710	SCLAV_0627	Putative glycosyl hydrolase
839217	357	SCLAV_0639	Putative membrane protein
894650	592	SCLAV_0692	Inositol-5-monophosphate dehydrogenase
939482	802	SCLAV_0735	Hypothetical protein
1232677	858	SCLAV_0990	Glycosyl transferase
1251940	705	SCLAV_1005	Phospholipid-binding protein
1469195	584	SCLAV_1210	Putative AraC-family transcriptional regulator
1522705	846	SCLAV_1255	Imidazole glycerol phosphate synthase subunit hisF
1560199	736	SCLAV_1289	DUF552 domain-containing protein
1560201	730	SCLAV_1289	DUF552 domain-containing protein
1648315	754	SCLAV_1366	Cytochrome C heme-binding subunit
1810740	885	SCLAV_1511	putative dehydrogenase
1945192	1194	SCLAV_1642	putative DNA-binding protein
2229288	615	SCLAV_1885	Transcriptional regulator, MarR family
2345328	860	SCLAV_1974	Glycosyl Hydrolase family 18 protein
2352618	622	SCLAV_1980	Putative polar amino acid ABC transporter permease protein
2562043	836	SCLAV_2152	Hypothetical protein
2676599	902	SCLAV_2259	Secreted protein
2676973	729	SCLAV_2259	Secreted protein
2750870	870	SCLAV_2321	Penicillin acylase
2798531	981	SCLAV_2360	Pyrroline-5-carboxylate reductase
2931836	606	SCLAV_2497	DnaB domain protein helicase domain protein
3028194	949	SCLAV_2578	putative transcriptional regulator
3207610	873	SCLAV_2717	putative transporter
3236675	877	SCLAV_2741	CRISPR-associated protein Cas1
3348963	483	SCLAV_2840	Putative prephenate dehydratase
3398176	1195	SCLAV_2886	Integral membrane protein
3586290	896	SCLAV_3046	putative secreted protein
3586291	873	SCLAV_3046	putative secreted protein
3645166	1066	SCLAV_3096	Putative ABC transporter ATP-binding protein
3712105	1146	SCLAV_3152	ATP/GTP-binding protein
3914298	1010	SCLAV_3349	Putative ABC transport system ATP-binding protein
4078185	758	SCLAV_3497	large ATP-binding protein
4244729	1052	SCLAV_3643	30S ribosomal protein S3
4387536	723	SCLAV_3777	Hypothetical protein
4506929	9	SCLAV_3886b	putative glycosyl transferase
4795383	644	SCLAV_4138	N,O-diacetyl muramidase
4807732	817	SCLAV_4150	Hypothetical protein
4885031	805	SCLAV_4205	3'-hydroxymethylcephem-O-carbamoyltransferase
5143196	699	SCLAV_4429	Acetyltransferase
5143197	698	SCLAV_4429	Acetyltransferase
5287640	698	SCLAV_4545	50S ribosomal protein L19
5328161	477	SCLAV_4582	Hypothetical protein
5382981	795	SCLAV_4623	Putative FtsK/SpoIIIE family protein
5858605	823	SCLAV_4978	subtilisin-like protease
5985971	702	SCLAV_5095	DNA-binding protein
6211944	873	SCLAV_5271	Non-ribosomal peptide synthetase
6360266	507	SCLAV_5409	ArsR family transcriptional regulator
6360267	512	SCLAV_5409	ArsR family transcriptional regulator
6396333	547	SCLAV_5443	Transcriptional regulator
6544565	770	SCLAV_5549	Cytochrome P450

B.2 Mutations in non coding regions which are upstream of one coding sequence.

Table B.3. Mutations in plasmid's intergenic regions which are upstream of one coding sequence.

Position	Reads	Downstream locus	Annotation
155100	5	SCLAV_p0143	Hypothetical protein
431607	172	SCLAV_p0446	Hypothetical protein
560480	4	SCLAV_p0555	Hypothetical protein
1307226	46	SCLAV_p1183	Thioesterase
1307227	39	SCLAV_p1183	Thioesterase
1651498	3	SCLAV_p1434	Hypothetical protein
1651499	3	SCLAV_p1434	Hypothetical protein
1769722	18	SCLAV_p1569	Secreted protein

Table B.4. Mutations in chromosomal intergenic regions which are upstream of one coding sequence.

Position	Reads	Downstream locus	Annotation
579023	15	SCLAV_0428	TetR-type regulator
579025	5	SCLAV_0428	TetR-type regulator
3318271	692	SCLAV_2814	Peptidoglycan-binding domain 1 protein
3657310	761	SCLAV_3107	Hypothetical protein
3807163	967	SCLAV_3244	Hypothetical protein
3990916	644	SCLAV_3423	Hypothetical protein
4000392	918	SCLAV_3431	Chitin-binding protein
4077207	639	SCLAV_3497	Large ATP-binding protein
4882915	323	SCLAV_4204	Positive regulator
5158988	41	SCLAV_4441	Hypothetical protein
5158990	26	SCLAV_4441	Hypothetical protein
5919761	5	SCLAV_5034	Putative secreted protein

B.3 Mutations detected in both strains when mapped to the DSM reference

Table B.5. Mutations detected in the plasmid in both strains when mapped to the DSM reference.

Position	Base in DSM reference	Base in WT and mutant strain	Reads in WT strain	Reads in Mutant strain	Total reads (both strains)	Locus
219408	A	G	131	29	160	intergenic
219415	T	G	99	32	131	intergenic
219425	C	G	78	30	108	intergenic
808943	G	C	109	120	229	SCLAV_p0763
827227	C	G	21	4	25	intergenic
850490	A	G	46	49	95	intergenic
862126	G	T	96	124	220	intergenic
1094556	T	G	25	28	53	SCLAV_p1022
1156882	T	C	3	6	9	intergenic
1158324	A	T	16	19	35	intergenic
1252930	A	C	4	3	7	intergenic
1252933	T	G	18	26	44	intergenic
1643795	C	A	252	21	273	intergenic

Table B.6. Mutations detected in the chromosome in both strains when compared to the DSM reference

Position	Base in DSM reference	Base in WT and mutant strain	Reads in WT strain	Reads in Mutant strain	Total reads (both strains)	Locus
85521	C	G	7	19	26	SCLAV_0028
182053	T	C	6	18	24	Intragenic
182054	T	G	9	28	37	Intragenic
311597	A	C	9	15	24	Intragenic
631695	G	C	7	18	25	Intragenic
631698	A	C	20	39	59	Intragenic
708353	C	G	74	42	116	SCLAV_0529
708469	T	C	8	20	28	Intragenic
708470	C	A	12	20	32	Intragenic
804381	C	G	256	333	589	SCLAV_0619
815806	T	C	126	132	258	Intragenic
815819	C	G	25	42	67	Intragenic
1199184	A	G	32	78	110	SCLAV_0962
1504887	T	C	36	56	92	SCLAV_1236b
1504895	T	C	59	85	144	SCLAV_1236b
1512350	G	C	7	24	31	Intragenic
1577425	T	G	229	290	519	Intragenic
1577426	T	G	233	304	537	Intragenic
1589556	G	A	246	292	538	Intragenic
1589559	A	G	246	295	541	Intragenic
1716518	T	G	346	442	788	Intragenic
1784768	T	C	156	107	263	SCLAV_1486
1812203	A	C	15	12	27	Intragenic
1937836	T	C	14	23	37	Intragenic
2021016	A	C	21	23	44	Intragenic
2111544	A	C	201	304	505	Intragenic
2151191	G	C	11	26	37	Intragenic
2151298	C	G	7	7	14	Intragenic
2268313	T	G	9	11	20	Intragenic
2293663	A	T	144	217	361	Intragenic
2396035	C	A	28	25	53	Intragenic
2592051	T	C	248	321	569	Intragenic
2592057	A	G	200	241	441	Intragenic
2592081	A	C	44	51	95	Intragenic
2874929	T	G	267	185	452	Intragenic
2874933	A	G	244	167	411	Intragenic
2874954	G	C	314	293	607	Intragenic
3084667	G	C	23	43	66	Intragenic
3084669	A	G	4	25	29	Intragenic
3086868	G	C	4	3	7	Intragenic
3129145	T	C	4	5	9	Intragenic
3181221	T	G	92	173	265	Intragenic
3243805	A	G	122	98	220	Intragenic
3243823	G	C	286	258	544	Intragenic
3405727	C	T	45	82	127	Intragenic

Position	Base in DSM reference	Base in WT and mutant strain	Reads in WT strain	Reads in Mutant strain	Total reads (both strains)	Locus
3405728	G	C	41	63	104	Intragenic
3405753	T	C	71	86	157	Intragenic
3405755	T	C	98	95	193	Intragenic
3634508	C	G	9	24	33	SCLAV_3088
3673421	C	G	226	431	657	SCLAV_3123
3673469	A	G	391	528	919	SCLAV_3123
4199151	A	G	152	146	298	SCLAV_3600
4246538	T	G	55	39	94	Intragenic
4284866	T	C	194	223	417	SCLAV_3691
4301301	C	G	117	168	285	Intragenic
4340908	T	G	7	18	25	Intragenic
4461849	G	C	4	4	8	Intragenic
4529943	A	G	291	244	535	Intragenic
4552493	C	G	16	39	55	Intragenic
4569499	A	C	284	310	594	Intragenic
4577595	T	C	9	3	12	Intragenic
4629273	A	G	141	154	295	Intragenic
4725674	T	G	9	8	17	SCLAV_4074b
4738241	C	G	628	725	1353	SCLAV_4087
4774344	G	C	499	560	1059	SCLAV_4124
4836388	C	G	14	31	45	Intragenic
4836389	T	G	16	52	68	Intragenic
4836402	C	A	145	168	313	Intragenic
5007135	T	C	112	104	216	Intragenic
5077649	T	G	587	585	1172	SCLAV_4372
5097346	G	C	472	709	1181	Intragenic
5231882	G	C	10	29	39	SCLAV_4494
5336965	A	C	281	342	623	SCLAV_4587
5336983	T	C	263	327	590	SCLAV_4587
5497447	A	C	122	229	351	Intragenic
5497592	C	G	24	52	76	SCLAV_4707
5497593	G	C	37	59	96	SCLAV_4707
5581312	A	C	6	13	19	Intragenic
5605655	T	G	428	622	1050	SCLAV_4775
5621170	G	C	15	7	22	Intragenic
5645937	A	G	1121	994	2115	Intragenic
6085182	T	G	113	198	311	Intragenic
6085208	A	G	95	121	216	Intragenic
6085231	C	A	15	22	37	Intragenic
6176463	C	A	240	274	514	Intragenic
6193892	A	C	370	454	824	SCLAV_5255
6194003	G	T	444	526	970	SCLAV_5255
6304193	T	C	6	14	20	Intragenic
6433662	A	G	5	12	17	SCLAV_5475
6727059	T	G	57	38	95	Intragenic

Appendix C
Discerning key parameters influencing high productivity and quality through recognition of patterns in process data

C.1 Summary

Offline, online, and materials data from 51 monoclonal antibody production runs were used to identify patterns prominent in the deviation of final product titer and final quality as measured by glycosylation and ion exchange chromatography (IEC) using a support vector regression approach. The analysis consisted of two stages. In the first stage, process parameters were used to construct a model to predict the process outcome (objective function). Models for three objective functions (final titer, Gal0, and IEC's acidic peak) were constructed. In the second stage, the impact of different parameters to the objective function was determined. Microarray data was used to complement this analysis.

C.2 Introduction

Monoclonal antibodies with clinical use are produced by recombinant DNA technology using mostly Chinese hamster ovary (CHO) or murine cell lines (Birch and Racher 2006). In commercial process development it is important to increase product throughput while maintaining product quality, and titers in the range 5-10 g/L have been reported in fed-batch cultures. A typical cell culture process includes cell expansion in several seed trains, with each subsequent vessel having increased capacity. The culture systems in common use in commercial production are fed-batch and continuous perfusion culture. In fed-batch systems, by far the most common, the production bioreactor can reach scales of up to 20,000 L. The feed includes nutrients that help maintain a level appropriate for growth.

Process optimization includes several parts of the process. For example, cells have been engineered to have extended life time by shutting down apoptosis or to avoid ammonia and lactic acid formation, which are toxic to cell growth (Jain and Kumar 2008). Media optimization and its feed is another part of the process in which optimization has focused.

Among the parameters that are controlled are pH, dissolved oxygen concentration, temperature, and osmolarity. Electronic records are typically kept for physical and process parameters, as well as materials used in the process. The measurements can be online, and have very high frequencies, or offline (like viable cell density) and have lower and non-uniform frequencies. However, control of biological process is complex and final product titer for multiple runs can have a wide distribution.

In 2004, the FDA published the guidance Process Analytical Technology (PAT) for Biopharmaceutical Products. The goal in PAT is to ensure final product quality, by understanding the process and combining monitoring of raw material and in-process attributes in real-time (Read, Park et al. 2010). The implementation of PAT requires process analyzers, as well as process control tools (Read, Park et al. 2010).

The quality by design (QbD) initiative has also received attention in the biopharmaceutical companies. Implementation of QbD requires understanding of the relationship between the critical quality attributes (CQAs) and the clinical properties of the product, as well as understanding of the relationship between the process and CQAs and the variability in raw materials (Rathore 2009).

PAT and QbD put an emphasis on developing processes that can produce high quality products by relying in automatic monitoring and control of critical parameters. These initiatives focus on a deep understanding of the process, coupled with control strategies in order to develop flexible and efficient processes. In order to achieve a deep understanding of the process, large datasets have to be investigated. Among the data mining techniques that have been used on bioprocess data are principal components regression (PCR), partial least squares (PLS), and artificial neural networks (ANN) (Teixeira, Oliveira et al. 2009).

C.2.1 Process and data overview

The data used in this analysis corresponds to 51 production runs at three scales: 10L, 100L, and 1000L. In the small scales (10L and 100L) cells were cultured for approximately 100 hours. At the 1000L scale the cells were maintained for roughly 350 hours (Figure C.1). For each run, data corresponding to process parameters measured online and offline was available. Lot number for some raw materials was also available.

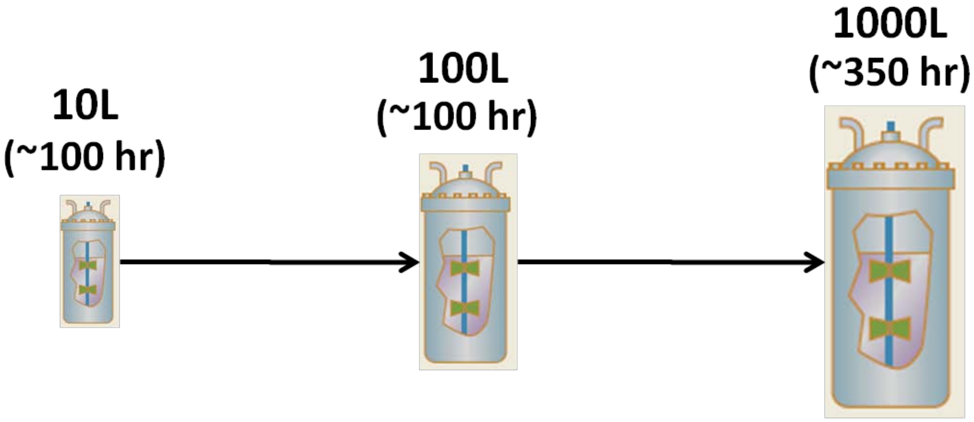


Figure C.1. Overview of monoclonal antibody production process.

Data for thirteen parameters measured offline was available for all scales. In addition, pH and osmolarity (also measured offline) were available for the 1000L scale only.

Table C.1. List of parameters used in this analysis

Offline	Online	Raw materials
Viable Cell Density (VCD) Viability Glucose Glutamine Lactate Ammonia LDH Volume CO2 Glutamate Sodium Potassium Titer pH (1000L only) Osmolarity (1000L only)	Pressure Stirrer speed pH Dissolved oxygen Temperature Aeration rate (100L and 1000L only) Filling volume (100L and 1000L only)	Soy hydrolysate reactor Soy hydrolysate feed Media reactor Media feed Pluronic reactor Rice hydrolysate feed (1000L only)

C.3 Methods

C.3.1 Offline data treatment

Data measured offline was not collected uniformly. Due to the non-uniform sampling intervals, the offline measurements were processed by linear interpolation. As can be seen in Figure C.2, the interpolated values closely resemble the original data points, indicating that linear interpolation adequately represents the dynamics of the data.

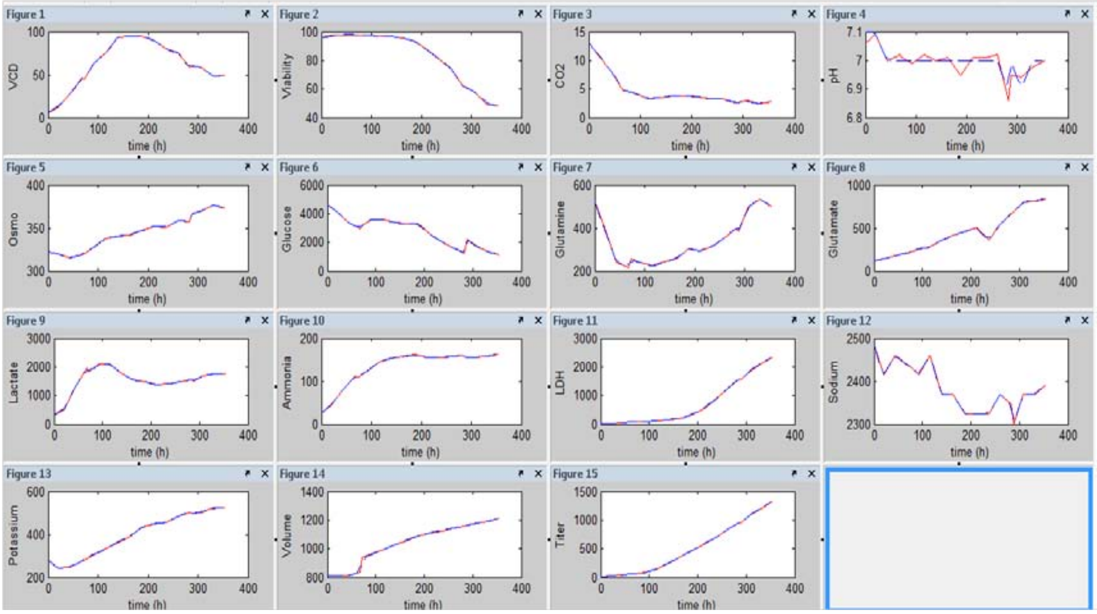


Figure C.2. Linear interpolation/extrapolation for offline data. Red curves: original data. Blue curves: interpolated data.

C.3.2 Online data treatment

Five parameters measured online (Table C.1) were available for all scales. In addition, aeration rate and filling volume were available for the 100L and 1000L scales only. Online parameters were uniformly measured every 10 minutes, thus linear interpolation/extrapolation was done only to fill missing values.

The online data was pre-processed using a moving window average (MWA) method to reduce the disturbances at the local time scales. At every time point *t*, a parameter's value was estimated as the average of all the measurements at the 10 subsequent time points (*t*, *t*+1, *t*+2, …, *t*+9). Figure C.3 shows the original profile (black curve) of stirrer speed at 100L scale for one run, and the resulting processed profile (red curve), which successfully delineates the temporal measurements.

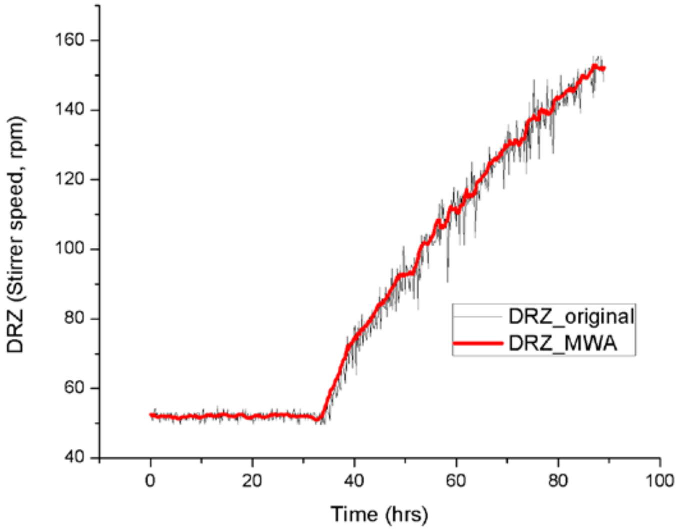


Figure C.3. Moving window average (MWA) for online data processing.

C.3.3 Raw materials data treatment

The identifiers for soy hydrolysate lot and media lot used in the bioreactor and the feed as well as the lot of pluronic used in the bioreactor were available for all three scales. For the 1000L bioreactor the identifiers for rice hydrolysate lot used in the feed were also available.

In most of the runs multiple lots of raw materials were used in each scale, however, the quantity of each lot used was not available. For such cases, all lots were treated as used in equal amounts.

C.3.4 Microarray data

A set of twenty-two microarray data previously analyzed in our lab (Kantardjieff 2009) correspond to cell samples taken from runs in this project. Quantile normalized data was used for this analysis. Table C.2 indicates the campaign (run), scale, and day (d) for which samples were taken for microarray data.

Table C.2. Microarray samples from current bioprocess data.

Campaign	10L	100L	1000L					
	d3	d3	d3	d6	d8	d10	d13	d15
D037.01E	Y	Y	Y	Y	Y	Y	Y	Y
D038.01E				Y			Y	
D037.02E	Y	Y	Y		Y	Y		Y
D038.05E	Y	Y	Y		Y	Y		Y

Differential expression detection was done using Significance Analysis of Microarrays (SAM) (Tusher, Tibshirani et al. 2001) and linear models (Smyth 2004) in the R programming language version 2.10.1 with the packages *siggenes* version 1.22.0 and *limma* version 3.4.4.

C.3.5 Estimation of similarity between any two runs

The likeness of any two runs was calculated in two steps as described in (Charaniya, Le et al. 2010). Briefly, in the first step each parameter is compared between any two runs using Euclidean distance. The calculated Euclidean distances were scaled between 0 and 1 and converted to a similarity measure by using an exponential transformation. For each parameter, a matrix containing the similarity calculations for all possible pairwise combinations was obtained. The resulting matrix is symmetric. Due to missing data in some runs, the matrix had different dimensions for the three objective functions tested: 51×51 for final titer, 40×40 for Gal0, and 46×46 for IEC's acidic peak. For raw material lots, an *n*-dimensional vector was constructed; *n* is the total number of different lots used. When a lot was used in a particular run its value is one, and when not used is zero. When multiple lots were used in the same run their contribution was recorded as equal between all the lots used. In the second step, all individual similarity matrices were integrated to obtain the overall similarity matrix for all pairwise comparisons of runs. In particular, the overall similarity score between any two runs is the weighted combination of the similarity scores for all parameters between these two runs.

C.3.6 Estimation of parameter weight

A weight was assigned to each parameter in proportion to its contribution to the deviation in process outcome (either final titer, Gal0, or IEC acidic peak). For each parameter, the weight is the Spearman correlation coefficient between the similarity scores of that parameter for all possible pairwise combinations of runs and the differences in objective function (Charaniya, Le et al. 2010). The weights of all parameters are scaled such that they sum up to one. Following this weighting scheme, critical parameters have a higher contribution to the overall similarity between any two runs.

C.3.7 Model training and evaluation

To investigate the train progression, the process data obtained from the 51 runs (40 in the case of Gal0, and 46 in the case of IEC acidic peak) was organized into eight datasets (Figure C.4). The first dataset comprises the process data from the 10L bioreactor scale only. The second dataset includes the data from the 10L bioreactor and the 100L bioreactor. Thus each stage incorporates all the data prior to that point. The division of the 1000L scales into different days was based on the points for which microarray data was available.

C.3.8 Support Vector Regression (SVR)

Support Vector Regression, an extension of Support Vector Machines was used to predict the process outcome. Three variables were used as process outcome: final titer, Gal0 (glycosylation), and IEC's acidic peak. The last two are related to product quality. LIBSVM (Chang and Lin 2001), an implementation of SVR in C was used for training and validating the model. A 10-fold cross validation scheme was used to assess the constructed SVR models. Model predictability was evaluated with two criteria, Pearson correlation (*r*) and root mean square error (RMSE).

C.3.9 Clustering product quality data

Product quality data (glycosylation parameters Gal0, Gal1, Gal2, and NG) were clustered using K-means clustering with *k*=2 within Spotfire version 9.1.2. Spotfire was also used for visualization purposes.

C.4 Results

Offline, online, and raw materials data from 51 runs from scales 10L, 100L, and 1000L was investigated to identify parameters critical to bioprocess. The analysis consisted of two stages:

1. Construction of a model to predict process outcome, and

2. Identification of parameters critical to process outcome.

The data was organized into eight datasets (Figure C.4).

The process outcome was explored using three objective functions: final titer, Gal0, and IEC's acidic peak.

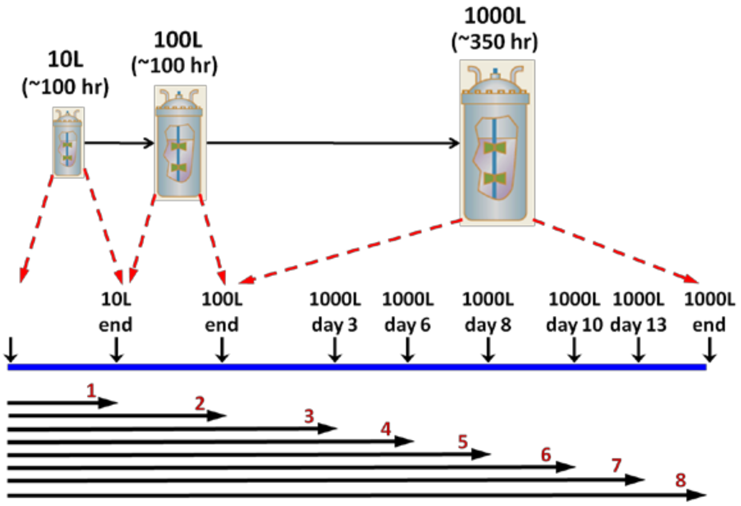


Figure C.4. Train progression divided into eight sets.

C.4.1 Stage 1: Model construction

C.4.1.1 Preliminary analysis for final titer indicates that small scales have low predictability

Final titer presents a wide distribution, spanning from 0.88 to 2.00 g/L (Figure C.5). The use of data from all three scales (10L, 100L, and 1000L) vs. use of data from the 1000L scale only was investigated in a preliminary analysis. The inclusion of all online, offline, and raw materials data was also explored in the preliminary analysis. Specifically the analysis was done with and without the inclusion of antibody titer ([mAb]) measured during the runs, as final titer is affected by its previous values. The inclusion of raw materials data was also explored, due to the lack of quantitative data for the multiple lots used during a single run.

In summary, four combinations were explored:

1. All offline and online parameters only, without including [mAb] as parameter. This case will be referred to as (offline + online).

2. All offline and online parameters, plus all raw materials, without including [mAb] as parameter. This case will be referred to as (offline + online + materials).
3. All offline and online parameters, including [mAb] as parameter. This case will be referred to as (off + on + [mAb]).
4. All offline and online parameters, including [mAb] and all raw materials. This case will be referred to as (off + on + lot + [mAb]).

The model predictability was assessed using Pearson correlation between actual vs. predicted final titer. From the plots in Figure C.6 it can be seen that predictability is low for the small scales (10L and 100L) when the data on materials was not included, that is, in the (offline + online) and (offline + online + [mAb]) combinations (black bars). When the raw materials data is included, predictability increases significantly for the small scales (Figure C.6 and Table C.3). This increase could be due to the same lots being used in the initial and final scale, being in reality the final scale the one which has the most information content for predictability. All four combinations were repeated using only information on the 1000L scale (blue bars in Figure C.6 and values in Table C.4). This analysis revealed that the inclusion of raw materials data results in an increased correlation in the initial stage (day0 – day 3), but a slightly lower correlation when all data was used (all days).

With or without materials data included in the analysis a marked improvement in correlation occurs between the datasets up to day 3 and up to day 6 of the 1000L scale. The correlation of datasets after day 6 of the 1000L scale continue to improve, but the change is gradual.

Due to the low predictability of the small scales and the fact that no quantitative data was available for materials, in the following sections only the results for the combination (online + offline + [mAb]) will be presented.

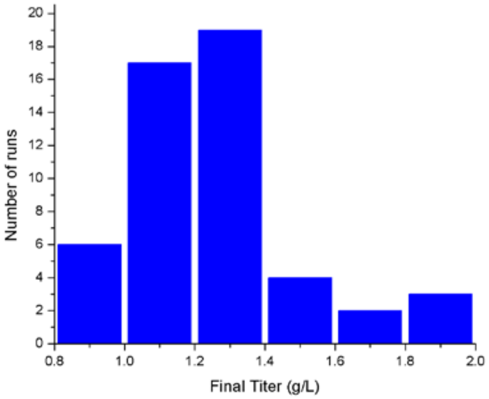


Figure C.5. Final titer distribution for all 51 runs.

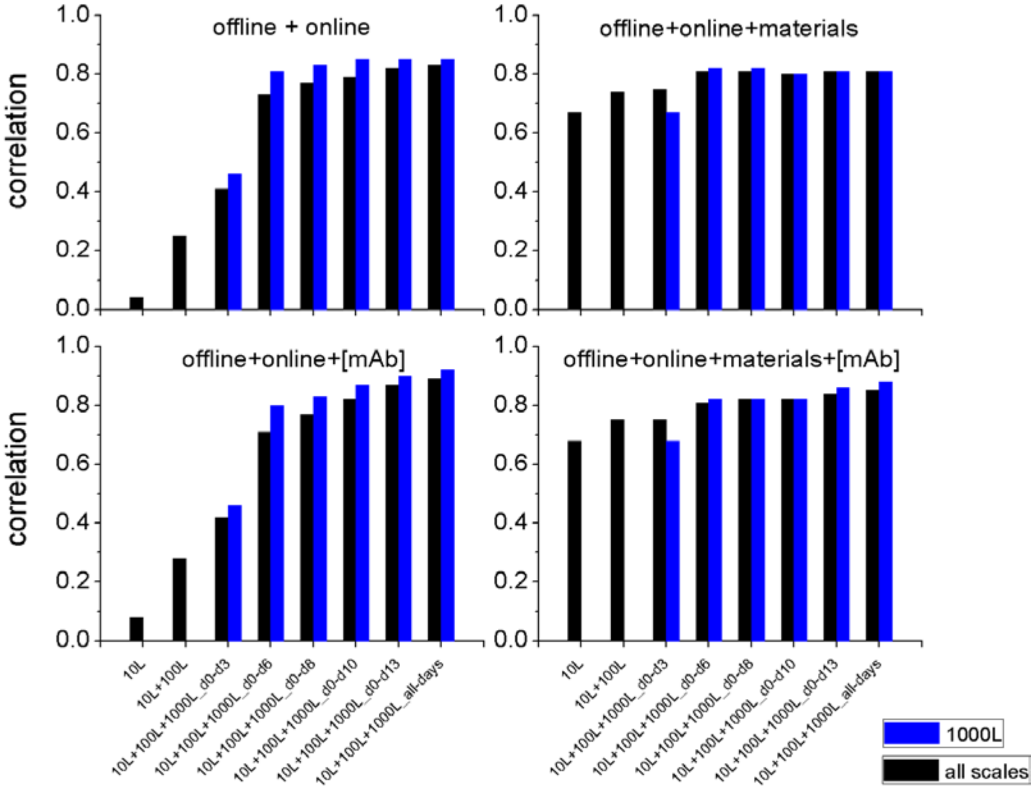


Figure C.6. Pearson correlation for different parameter combinations in the four cases tested.

Table C.3. Pearson correlation for different parameter combinations.

Scale	Dataset	offline + online	offline + online + materials	offline + online + [mAb]	offline + online + materials + [mAb]
10L	all days	0.04	0.67	0.08	0.68
10L+100L	all days	0.25	0.74	0.28	0.75
10L+100L+1000L	day0 – day3	0.41	0.75	0.42	0.75
10L+100L+1000L	day0 – day6	0.73	0.81	0.71	0.81
10L+100L+1000L	day0 – day8	0.77	0.81	0.77	0.82
10L+100L+1000L	day0 – day10	0.79	0.80	0.82	0.82
10L+100L+1000L	day0 – day13	0.82	0.81	0.87	0.84
10L+100L+1000L	all days	0.83	0.81	0.89	0.85

Table C.4. Pearson correlation for different parameter combinations using only 1000L scale data.

Scale	Dataset	offline + online	offline + online + materials	offline + online + [mAb]	offline + online + materials + [mAb]
1000L	day0 – day3	0.46	0.67	0.46	0.68
1000L	day0 – day6	0.81	0.82	0.80	0.82
1000L	day0 – day8	0.83	0.82	0.83	0.82
1000L	day0 – day10	0.85	0.80	0.87	0.82
1000L	day0 – day13	0.85	0.81	0.90	0.86
1000L	all days	0.85	0.81	0.92	0.89

C.4.1.2 Predictability for final titer is high

Due to low predictability of the small scales (10L and 100L) when materials data was not included, and the lack of quantitative data for the raw materials, the final analysis for final titer, which is presented next, was done using the 1000L data only, for the combination (offline + online + [mAb]).

The correlation (*r*) for actual vs. predicted final titer is shown in Figure C.7. For the first dataset, that is, from the beginning of the 1000L scale and up to day 3, *r* = 0.46. A significant improvement in correlation occurs when data up to day 6 was included (*r* = 0.80). Correlation continues to increase gradually with the addition of subsequent data. When data for all days, that is, up to the end of the run is used, correlation reaches 0.92.

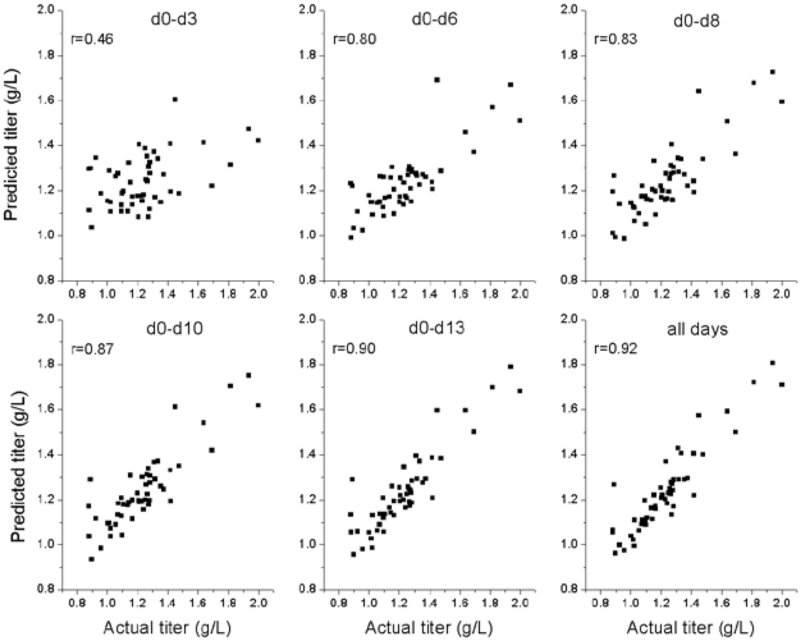


Figure C.7. Actual vs. predicted final titer for the combination (online + offline + [mAb]).

C.4.1.3 Predictability for Gal0 is high

Gal0 data was not available for all 51 runs, but only for 40. Gal0 distribution for the 40 runs covers the range from 49.1 to 67.8 (Figure C.8). The analysis for Gal0 was done using data from the 1000L scale only for the combination (offline + online + [mAb]).

The correlation (*r*) for actual vs. predicted Gal0 is shown in Figure C.9. For the first dataset, that is from the beginning of the 1000L scale and up to day 3, *r* = 0.61, higher than for final titer as objective function (*r* = 0.46). As in the case of final titer as objective function, there is a significant improvement in correlation when data up to day 6 was included (*r* = 0.85). A slight increase occurs between datasets up to day6 and up to day8 (*r* = 0.88) and then remains at that level (Table C.5 and Figure C.10).

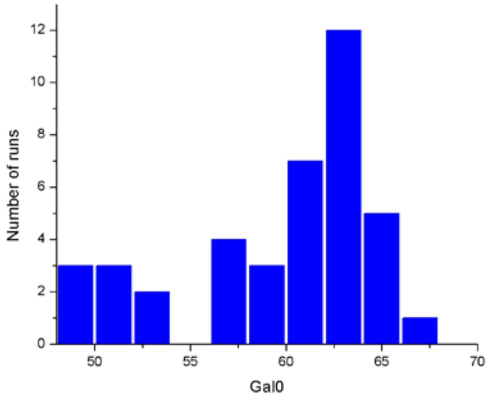


Figure C.8. Gal0 distribution for 40 runs.

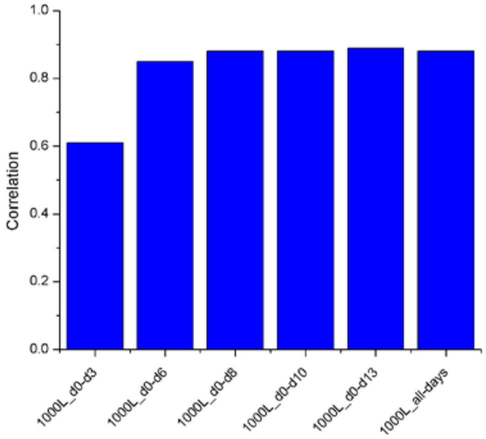


Figure C.9. Correlation for Gal0 for the combination (offline + online + [mAb]) using 1000L scale data.

Table C.5. Correlation for Gal0.

Scale	Dataset	offline + online + [mAb]
1000L	day0 – day3	0.61
1000L	day0 – day6	0.85
1000L	day0 – day8	0.88
1000L	day0 – day10	0.88
1000L	day0 – day13	0.89
1000L	all days	0.88

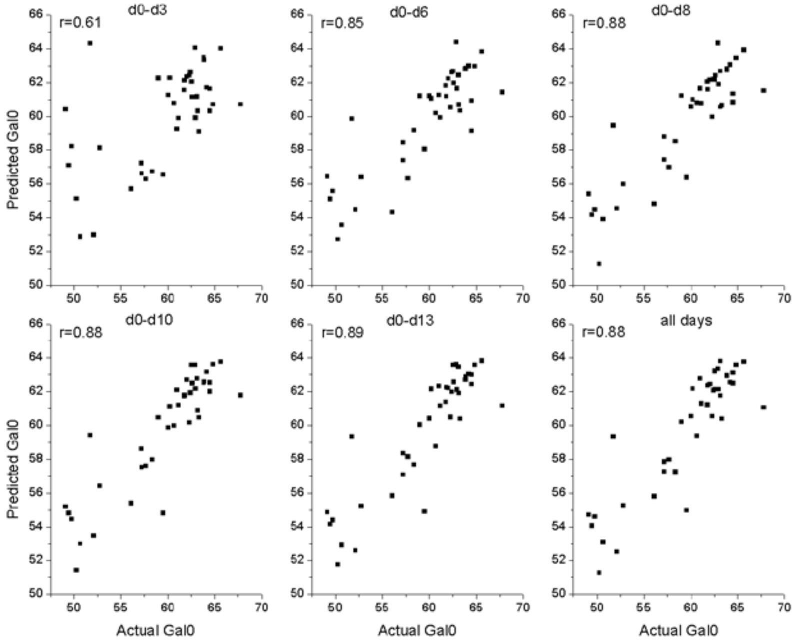


Figure C.10. Actual vs predicted Gal0 for (online + offline + [mAb] using 1000L data.

C.4.1.4 Predictability for IEC's acidic peak is low

IEC's acidic peak data was only available for 46 runs. The distribution for IEC's acidic peak values is narrow, with most data between 3.49 and 10.03 and only one run beyond this range (15.77) (Figure C.11). The analysis was only done for the 1000L scale for the combination (offline + online + [mAb].

Correlation between the prediced IEC's acidic peak values and the actual values is low. For the dataset up to day3 the correlation is 0.44 and increases to only 0.51 when all days are included in the analysis (Table C.6 and Figure C.12).

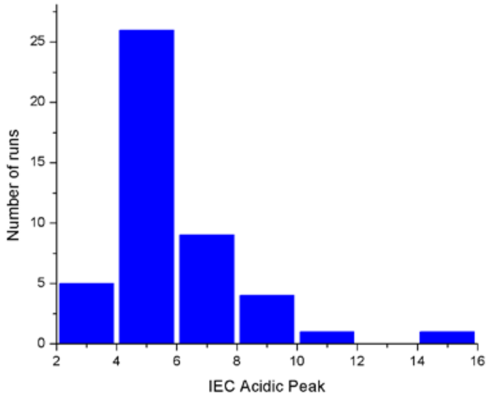


Figure C.11. IEC's acid peak distribution for 46 runs.

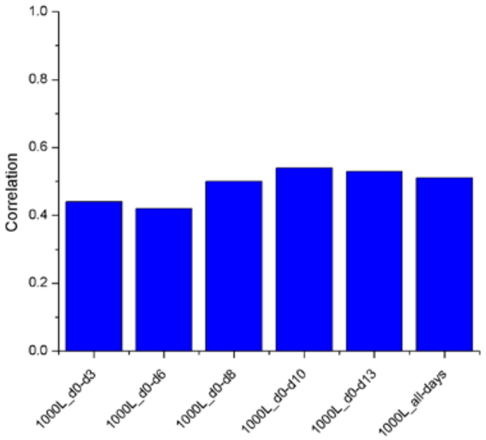


Figure C.12. Correlation for IEC's acidic peak for the combination (online + offline + [mAb]) using 1000L scale data.

Table C.6. Correlation for IEC's acidic peak.

Scale	Dataset	offline + online + [mAb]
1000L	day0 – day3	0.44
1000L	day0 – day6	0.42
1000L	day0 – day8	0.50
1000L	day0 – day10	0.54
1000L	day0 – day13	0.53
1000L	all days	0.51

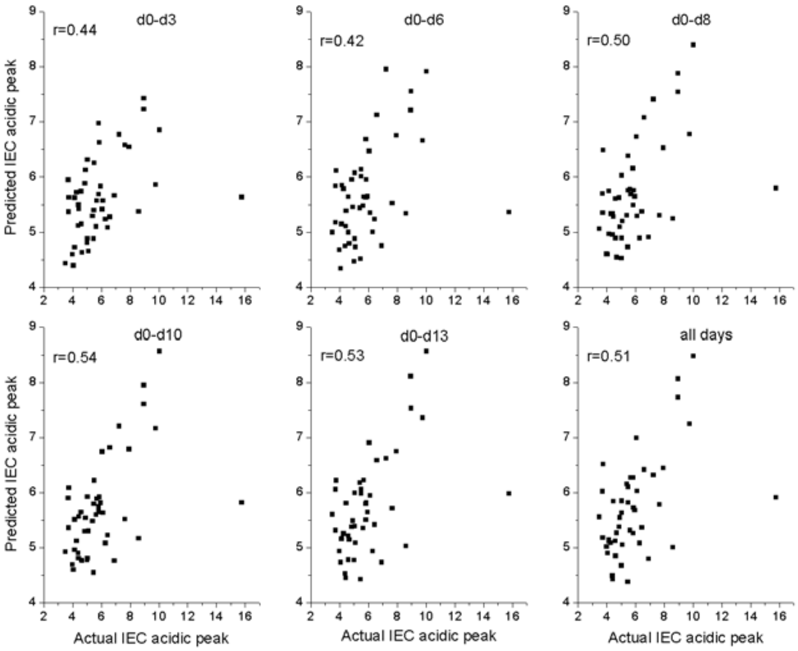


Figure C.13. Actual vs. predicted IEC's acidic peak for the combination (online + offline + [mAb]) using 1000L scale data.

C.4.2 Stage 2: Identification of critical parameters

In this analysis it was not only of interest to construct models that can predict process outcome, but also which parameters have an effect on it. As explained in

Section C.3.6 critical parameters have a higher contribution to the overall similarity between any two runs.

C.4.2.1 A few parameters have high impact on final titer

For the case of final titer as objective function and the combination (offline + online + [mAb]) a total of 22 parameters were considered. Figure C.14 shows the histogram of weights of 22 parameters. Most parameters have weights less than 0.2, implying that their contributions are not critical to the deviation in the process outcome. This value (0.2) was then used as threshold of weight to identify important parameters.

The parameter with the highest weight is [mAb], followed by stirrer speed, Viable Cell Density (VCD) and Glucose, and with lower weights LDH, ammonia, and lactate. Even though [mAb] has the highest weight, its impact manifests mostly at the end of the run, appearing as critical parameter only after data from day 10 was included in the analysis. In contrast, stirrer speed appears as a critical parameter in the very first dataset, which only includes information up to day 3 of the 1000L scale (Figure C.15). Table C.7 lists other parameters with high weights (> 0.2), even when they do not remain critical up to the end.

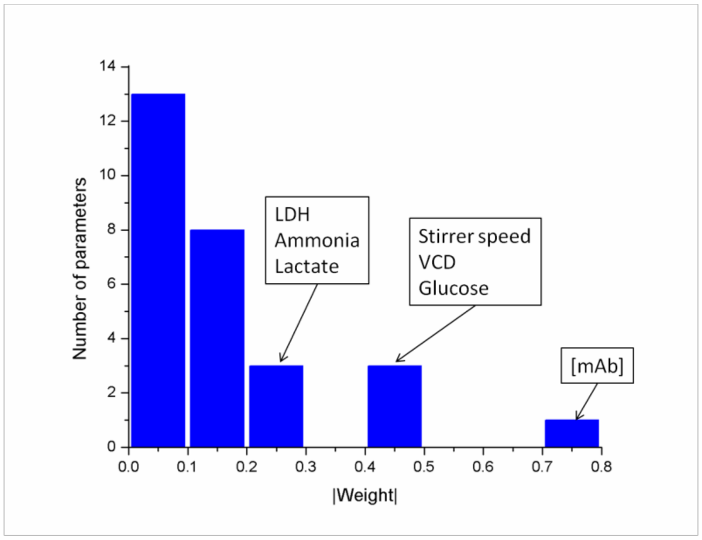


Figure C.14. Parameters with high impact on final titer.

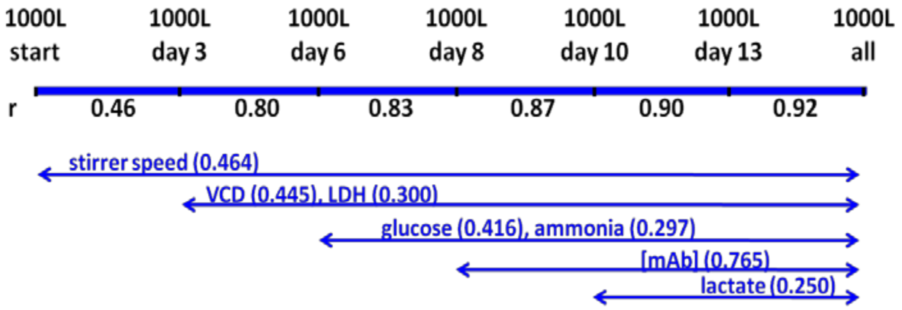


Figure C.15. Critical parameters for final titer manifest at different stages.

Table C.7. Weight for critical parameters in the different datasets for final titer.

Dataset	Parameter	Weight
d0-d3	PH	0.202
	DRZ	0.206

Dataset	Parameter	Weight
d0-d6	Glutamine	0.285
	LDH	0.296
	DRZ	0.311
	VCD	0.344

Dataset	Parameter	Weight
d0-d8	LDH	0.260
	Glucose	0.262
	Glutamine	0.295
	Ammonia	0.324
	DRZ	0.350
	VCD	0.362

Dataset	Parameter	Weight
d0-d10	PH	0.205
	Glutamine	0.284
	Titer	0.298
	LDH	0.302
	Ammonia	0.325
	Glucose	0.333
	DRZ	0.390
	VCD	0.401

Dataset	Parameter	Weight
d0-d13	Lactate	0.236
	Ammonia	0.279
	LDH	0.286
	Glucose	0.400
	VCD	0.430
	DRZ	0.444
	Titer	0.628

Dataset	Parameter	Weight
all days	Lactate	0.249
	Ammonia	0.297
	LDH	0.300
	Glucose	0.416
	VCD	0.445
	DRZ	0.464
	Titer	0.765

C.4.2.2 Parameters that impact final titer affect Gal0

For Gal0, 22 parameters were considered in the combination (offline + online + [mAb]). Figure C.16 shows the histogram of weights of all 22 parameters considered in the analysis. Since a weight threshold is not as clear as in the case of final titer, the same value of 0.2 was used.

Of the ten parameters with weight higher than the threshold of 0.2, seven of them are the same as the critical parameters detected in the case of final titer. The three additional parameters are CO_2, temperature, and viability. Stirrer speed and VCD are parameters with high weights and an early appearance as critical parameters (since dataset up to day3) (Figure C.17). Of the additional critical parameters CO_2 has an impact since day 3, whereas viability only appears as critical parameter at day 13. Table

C.8 lists other parameters with high weights (> 0.2), even when they do not remain critical up to the end.

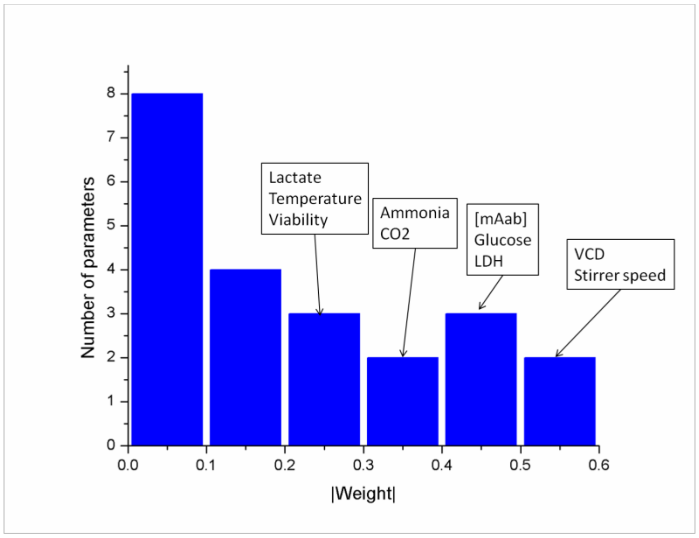


Figure C.16. Parameters with high impact on Gal0.

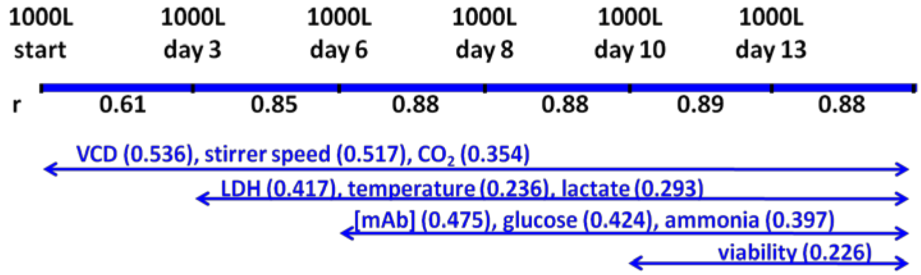

1000L
start
1000L
day 3
1000L
day 6
1000L
day 8
1000L
day 10
1000L
day 13
r
0.61
0.85
0.88
0.88
0.89
0.88
VCD (0.536), stirrer speed (0.517), CO₂ (0.354)
LDH (0.417), temperature (0.236), lactate (0.293)
[mAb] (0.475), glucose (0.424), ammonia (0.397)
viability (0.226)

Figure C.17. Critical parameters for Gal0 manifest at different stages.

Table C.8. Weight for critical parameters in the different datasets for Gal0.

Dataset	Parameter	Weight
d0-d3	CO2	0.222
	Lactate	0.263
	DRZ	0.265
	PH	0.270
	VCD	0.338

Dataset	Parameter	Weight
d0-d6	Sodium	0.214
	CO2	0.215
	LDH	0.217
	TEF	0.249
	PH	0.259
	Glutamine	0.295
	DRZ	0.427
	VCD	0.505

Dataset	Parameter	Weight
d0-d8	Titer	0.206
	Lactate	0.221
	Sodium	0.230
	TEF	0.246
	CO2	0.259
	PH	0.265
	LDH	0.278
	Glutamine	0.326
	Glucose	0.351
	Ammonia	0.361
	DRZ	0.453
	VCD	0.498

Dataset	Parameter	Weight
d0-d10	TEF	0.245
	Sodium	0.254
	Glutamine	0.275
	CO2	0.288
	Lactate	0.298
	Titer	0.302
	Glucose	0.387
	Ammonia	0.409
	LDH	0.441
	DRZ	0.477
	VCD	0.502

Dataset	Parameter	Weight
d0-d13	Viability	0.219
	TEF	0.241
	Lactate	0.318
	CO2	0.339
	LDH	0.398
	Ammonia	0.399
	Glucose	0.402
	Titer	0.450
	DRZ	0.511
	VCD	0.527

Dataset	Parameter	Weight
all days	Viability	0.226
	TEF	0.236
	Lactate	0.293
	CO2	0.354
	Ammonia	0.397
	LDH	0.417
	Glucose	0.424
	Titer	0.475
	DRZ	0.517
	VCD	0.536

C.4.2.3 Some critical parameters impacting IEC's acidic peak also impact Gal0 and titer

In the case of IEC's acidic peak as objective function, the weight distribution for the 22 parameters involved in the analysis does not present a clear cutoff (Figure C.18). Using the same threshold as that used for final titer and Gal0 there are seven parameters with weight higher than 0.2. Six of those parameters are common to critical parameters determined using final titer and Gal0 as process outcome. The only additional parameter is filling volume.

[mAb] and stirrer speed have the highest weights and they both manifest at the same time (dataset day0-day6). In this analysis, no "late" critical parameters are observed (Figure C.19). Critical parameters appear only in datasets day0-day6 and day0-day8 and continue for the rest of the run. Table C.9 lists other parameters with high weights (> 0.2), even when they do not remain critical up to the end of the run.

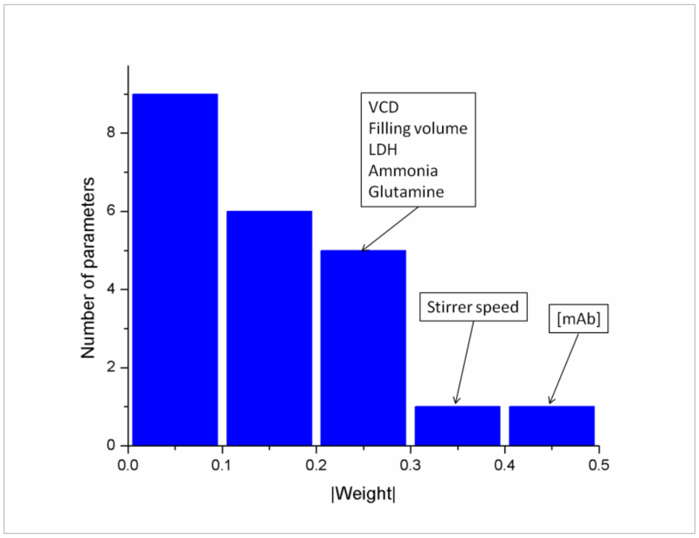


Figure C.18. Parameters with high impact on IEC's acidic peak.

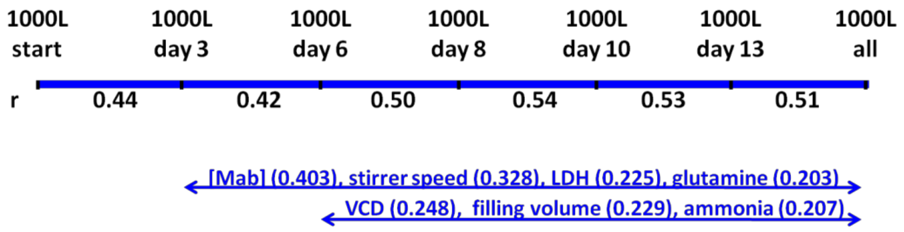

1000L start
1000L day 3
1000L day 6
1000L day 8
1000L day 10
1000L day 13
1000L all
r
0.44
0.42
0.50
0.54
0.53
0.51
[Mab] (0.403), stirrer speed (0.328), LDH (0.225), glutamine (0.203)
VCD (0.248), filling volume (0.229), ammonia (0.207)

Figure C.19. Critical parameters for IEC's acidic peak manifest at different stages.

Table C.9. Weight for critical parameters in the different sets for IEC acidic peak as process outcome.

Dataset	Parameter	Weight
d0-d3	TEF	0.282

Dataset	Parameter	Weight
d0-d6	Glutamine	0.248
	DRZ	0.265
	Titer	0.282
	LDH	0.283

Dataset	Parameter	Weight
d0-d8	FEF	0.202
	Glucose	0.204
	Glutamine	0.230
	VCD	0.234
	Ammonia	0.254
	Titer	0.266
	LDH	0.288
	DRZ	0.316

Dataset	Parameter	Weight
d0-d10	FEF	0.211
	Glutamine	0.215
	Glucose	0.217
	Ammonia	0.218
	LDH	0.237
	VCD	0.258
	Titer	0.281
	DRZ	0.317

Dataset	Parameter	Weight
d0-d13	Glucose	0.201
	LDH	0.204
	Ammonia	0.209
	Glutamine	0.223
	FEF	0.224
	VCD	0.251
	DRZ	0.321
	Titer	0.380

Dataset	Parameter	Weight
all days	Glutamine	0.203
	Ammonia	0.207
	LDH	0.225
	FEF	0.229
	VCD	0.248
	DRZ	0.328
	Titer	0.403

C.4.3 Critical parameters correlation to final titer

Visualization of critical parameters is important for identifying correlations with respect to process outcome, as well as to gain insight on the dynamics and range of the parameter profile. Four parameters appear as critical in all three (final titer, Gal0, and IEC acidic peak) analyses. These parameters are: stirrer speed, VCD, LDH, and ammonia. These parameters appear in Figure C.20. The plots are color coded according to final titer.

It is clear that a correlation exists between the critical parameters and final titer. A high final titer correlates to high stirrer speed, high VCD, and high LDH. A correlation is not clear between ammonia and final titer. In addition, two more parameters, glucose, and lactate, appear as critical in two (final titer, and Gal0) analyses (Figure C.21). Both critical parameters have an inverse correlation with final titer, that is, low glucose and low lactate correlate with high final titer. In the lactate plot, two runs with very high final lactate are seen. It is possible that the wide range of values for this parameter reduced the correlation with final titer, reducing its impact.

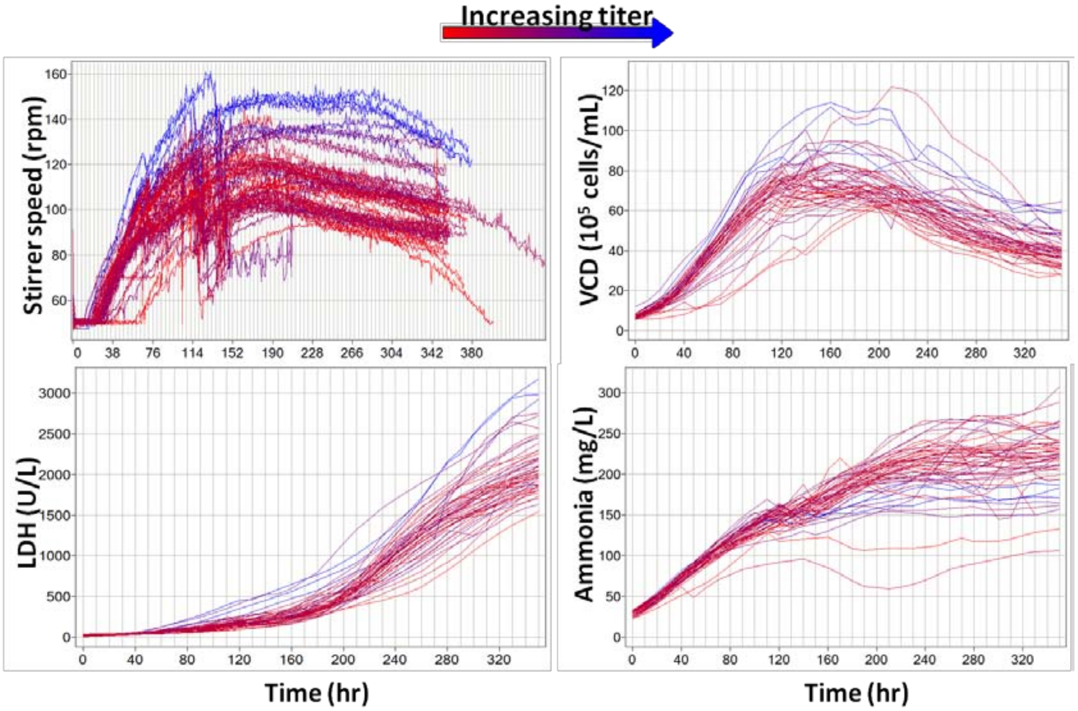


Figure C.20. Critical parameters in all three analyses.

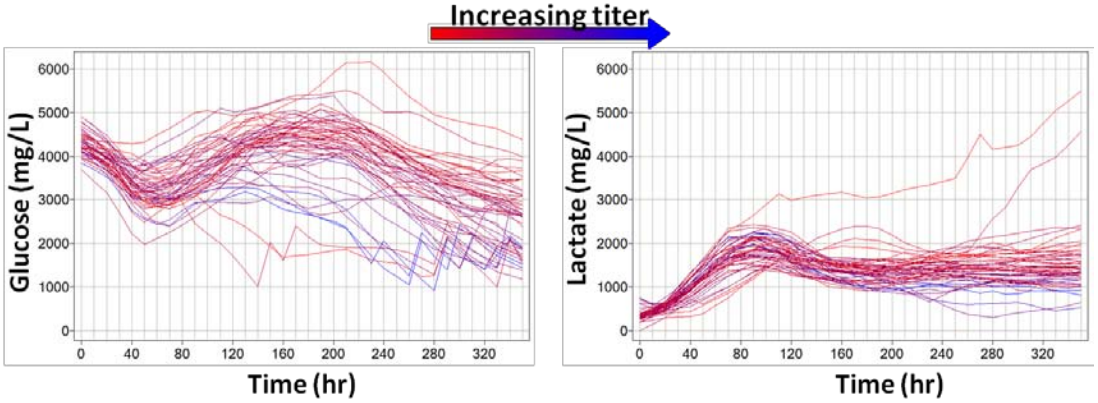


Figure C.21. Critical parameters in Final titer and Gal0 analyses.

C.4.4 Correlation between product quantity and quality

Since several critical parameters appeared in the analysis using final titer and Gal0 as objective functions, the possible relation of these two was explored. K-means clustering was performed in all parameters determined for Glycosylation method 1 (i.e., Gal0, Gal1, Gal2, and NG), with *k*=2 (Figure C.22).

All three glycosylation types (Gal0, Gal1, and Gal2) separate into two well defined clusters according to its corresponding high and low values. In the case of non glycosylated values (NG), the resulting clusters still contain both low and high values for NG. A 3D plot (Figure C.23) using Gal0, Gal1, and Gal2 as axis clearly separates the data into two defined clusters. Thus the possible correlation of the resulting clusters to final titer was explored. As seen in Figure C.24, there is a clear correlation between the glycosylation-based clusters and final titer. Cluster 1 mostly corresponds to high final titer, whereas cluster 2 corresponds to low final titer.

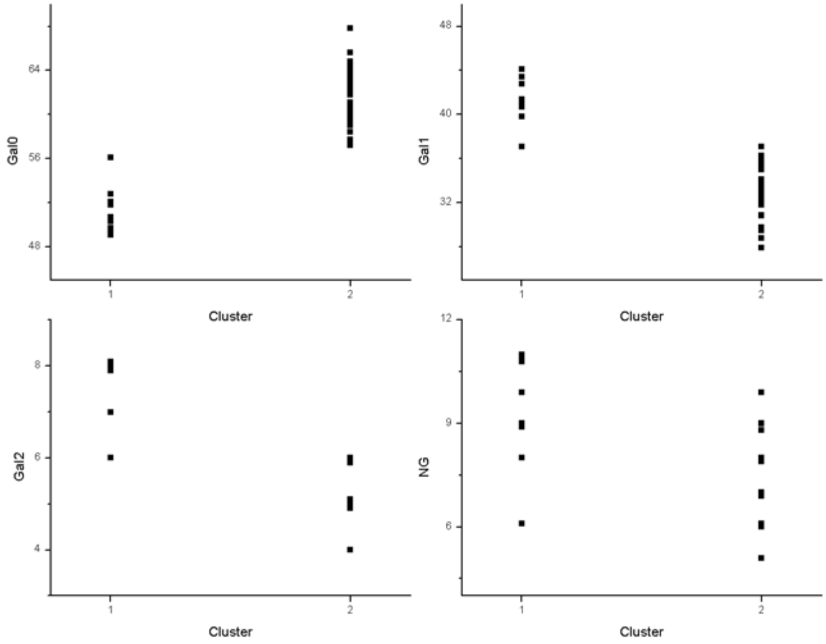


Figure C.22. K-means clustering (k=2) on glycosylation values.

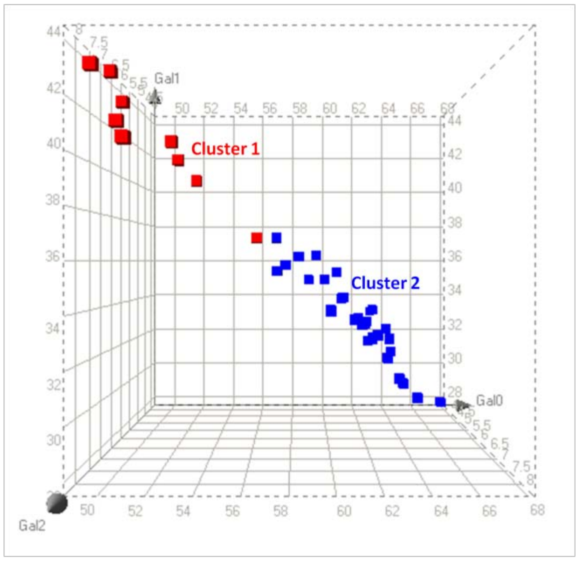


Figure C.23. 3D plot of Gal0, Gal1, and Gal2.

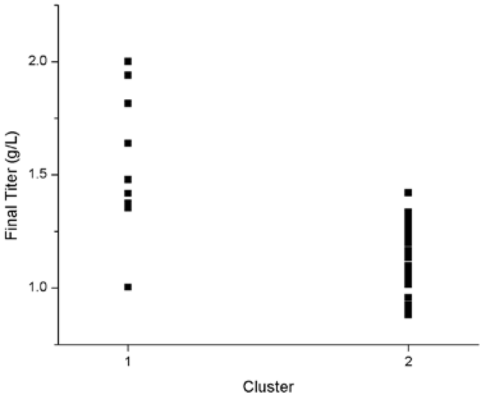


Figure C.24. Glycosylation-based clusters correlate with final titer.

C.4.5 Microarray data analysis

From the twenty-two microarray data corresponding to the 51 runs analyzed in this project, six of them are for the small scales, 10L and 100L. These arrays were not analyzed as the predictability of the models was low in these scales. The remaining 16 arrays, do correspond to both high and low titer (Table C.10). However, the possible comparisons are reduced to a comparison of one run with high titer (run 17) vs. two runs with low titer (runs 20 and 21), since the time points for the remaining run (run 18) do not overlap with the low titer runs. Only four time points exist in common between runs 17, 20, and 21. These time points are those for samples taken on days 3, 8, 10, and 15. Although the highest increase in predictability for final titer and Gal0 occurred between datasets day0-day3 and day0-day6, this comparison can't be explored for differences in titer since only run 17 was probed at both days.

The comparison (H.8-H.3)-(L.8-L.3), where H corresponds to high titer and L to low titer, and 8 and 3 indicate the day at which the sample was taken resulted in no differentially expressed genes using the microarray analysis software *limma*.

Table C.10. Array samples and corresponding final titer and glycosylation information.

Campaign	Run (1000L scale)	Glycosylation	Final Titer	1000L					
				d3	d6	d8	d10	d13	d15
D037.01E	17	cluster 1	high	Y	Y	Y	Y	Y	Y
D038.01E	18	cluster 1	high		Y			Y	
D037.02E	20	no data	low	Y		Y	Y		Y
D038.05E	21	cluster 2	low	Y		Y	Y		Y

C.5 Discussion

For the case of final titer as objective function, the analysis indicated that the small scales (10L and 100L) did not contribute significantly to model predictability. Predictability was assessed by the correlation of actual final titer vs. predicted final titer. Thus for the cases of Gal0 and IEC's acidic peak as objective functions, models were constructed using data of the 1000L scale only.

Predictability was high for final titer (0.92) and Gal0 (0.88), especially considering the sample size. The correlation for IEC's acidic peak, however, was low (0.51). The low predictability for the IEC's acidic peak case could be due to the narrow range of values in that parameter. Predictability increased substantially when data from day 0 to day 6 was analyzed as compared to that of the analysis using only data from day 0 to day 3. Predictability continued to increase with further data inclusion, but at a slower rate.

Stirrer speed, VCD, LDH, and ammonia were identified as critical parameters in all three (final titer, Gal0, and IEC's acidic peak) cases. The impact of stirrer speed and VCD was manifest from the early days. Glucose and lactate were identified as critical parameters for the final titer, and the Gal0 cases. No raw materials were identified as critical, probably due to the lack of quantitative data. In addition to shared critical parameters, a correlation was detected between glycosylation values and final titer.

Microarray data was used to complement this analysis. Comparison of days 3 vs. 6 was of particular interest due to the increase in model predictability. Unfortunately arrays of samples taken at these days were available for only one run, thus analysis to compare differences in final titer was not possible. Microarray data for days 3 vs. 8 was available for one run with high final titer and two runs with low final titer. No genes were identified as differentially expressed between the two days for the high titer vs. low titer runs. The reduced sample size could explain the lack of power to detect differences in gene expression.

The analysis of data from 51 runs has revealed an array of correlated parameters linked to final titer. These critical parameters are also important for product quality. Furthermore, there is a correlation between final titer and product quality, as measured by glycosylation profiles. Stirrer speed was in both cases the controllable parameter with highest weight, pointing to parameters that can be used for intervention.

C.6 Concluding remarks

This analysis reveals that the algorithm controlling stirrer speed contributes to differences in final outcome and thus points to the possibility of intervening to implement corrective actions. This analysis used process parameters only. Further correlations could be identified if calculated parameters were analyzed. Further applications of mining tools to larger data sets could facilitate the identification of corrective actions.