Genomics and Transcriptomics of the Industrial Acarbose Producer *Actinoplanes* sp. SE50/110

Ph. D. Thesis

submitted to the
Faculty of Technology,
Bielefeld University, Germany
for the degree of Dr. rer. nat.

by

Patrick Schwientek

January, 2012

Referees:

Prof. Dr. Alfred Pühler
Prof. Dr. Jens Stoye

Printed on non-aging paper according to DIN-ISO 9706.

Acknowledgments

This Ph. D. project was carried out in the period between January 2009 and January 2012 at the Center for Biotechnology (CeBiTec), Bielefeld University, Germany. It constitutes an interdisciplinary collaboration between the CeBiTec's Institute of Genome Research and Systems Biology, the CeBiTec's Institute of Bioinformatics, and the industrial partner Bayer HealthCare AG, Wuppertal, Germany. It is a pleasure to thank the many people who made this thesis possible.

First and foremost, I would like to express my deep and sincere gratitude to my supervisors Prof. Dr. Alfred Pühler and Prof. Dr. Jens Stoye for their stimulating guidance, invaluable support, and wide knowledge. It is due to their extensive experience and foresight that this Ph. D. project went so smoothly and will forever remind me of a challenging but also rewarding time. I am also deeply grateful to Dr. Jörn Kalinowski, for his productive advice and with whom I enjoyed many fruitful discussions. Furthermore, I wish to thank Prof. Dr. Karsten Niehaus for his expertise, help and patience during electron microscopy.

This thesis would not have been possible without the help and support of several experts from the wet-labs. I am especially thankful to Dr. Christian Rückert and Dr. Rapfael Szczepanowski who were involved in the preparation and supervision of the DNA-sequencing process. Moreover, I owe Sergej Wendler, Armin Neshat, Christina Eirich, and Katharina Pfeifer great thanks for the strain cultivation and preparation of the RNA-sequencing experiments. For enhancing my limited wet-lab competence and for her help and expertise with high-GC PCRs, I am very thankful to Yvonne Kutter. I also want to express my gratitude to Alexandra Tilker from the IIT GmbH for her efforts in testing specialized PCR protocols and for her repeated acceptance of sequencing orders after closing hour.

I am especially thankful to my cooperation partners from Bayer HealthCare AG, Dr. Klaus Selber and Dr. Andreas Klein for their kind support, interesting discussions, and delicious dinners. Furthermore, I wish to thank Dr. Bernd Weingärtner and Dr. Hermann Wehlmann for providing various information about *Actinoplanes* cultivations and for providing isolated DNA for genome sequencing. I gratefully acknowledge the funding by Bayer HealthCare AG that made my Ph. D. work possible.

I would like to thank the Cluster Industrial Biotechnology (CLIB^{2021}) for financial support and for the organization of many training courses and activities that I attended. I also want to thank my friends and fellow graduate students for making this Ph. D. an active, communicative, and very enjoyable time of my life.

In the end, I wish to thank my parents for their generous support in all my pursuits, especially during the last weeks of this thesis. Lastly, I would like to thank my partner Shokoufeh Ghezelbash for all her love and encouragement.

Abstract

Actinoplanes sp. SE50/110 is known as the wild type producer of the alpha-glucosidase inhibitor acarbose, a potent drug used worldwide in the treatment of type-2 diabetes mellitus. As the global incidence of diabetes is rapidly rising, an ever increasing demand for diabetes drugs, such as acarbose, needs to be anticipated. Consequently, derived *Actinoplanes* strains with increased acarbose yields are being used in large scale industrial batch fermentation, which were continuously optimized by mutagenesis and screening experiments. However, being applied for over 20 years, this conventional optimization strategy has now reached its limits and is generally superseded by modern genetic engineering approaches, which require the genome sequence of the organism.

Hence, the first part of this Ph. D. thesis dealt with the sequencing, assembly and annotation of the complete genome sequence of *Actinoplanes* sp. SE50/110, the first publicly available genome of the genus *Actinoplanes*. Due to its high GC-content of 71.32% and the formation of stable secondary structures that hindered the sequencing process, adapted protocols were developed which allowed the establishment of the complete sequence. The final genome consists of a single circular chromosome with a size of 9.4 Mb hosting about 8,400 genes. Besides the known acarbose biosynthetic gene cluster sequence, several new non-ribosomal peptide synthetase-, polyketide synthase- and hybrid-clusters were identified on the *Actinoplanes* genome. Another key finding represents the discovery of a functional actinomycete integrative and conjugative element, which might pose an elegant way of genetically accessing the organism. Phylogenetic analysis of the core genome revealed a rather distant relation to other sequenced species of the family Micromonosporaceae, whereas *Actinoplanes utahensis* was found to be the closest species based on 16S rDNA comparison.

The second part of this work complemented the genomic information with transcriptome experiments using RNA-sequencing technology. These analyses resulted in the discovery of non-coding RNAs, novel protein coding sequences, and antisense transcripts to known genes, which lead to an improved annotation of the *Actinoplanes* sp. SE50/110 genome. Moreover, genome wide expression quantification provided – for the first time – insights into the transcriptional landscape of the acarbose producer. In this regard, differential expression testing between three different *Actinoplanes* cultivations were also performed in order to elucidate the changes in gene expression in response to varying growth-media compositions. It was found that the different media had significant impact on growth rate and acarbose productivity, which was clearly reflected on the transcriptional level. In particular, the acarbose biosynthesis gene cluster happend to be highly up-regulated in maltose-containing media and almost silent in the glucose-containing medium. Additionally, one of the identified non-ribosomal peptide synthetase gene clusters showed high expression, which resembled the expressional pattern of the acarbose cluster across the analyzed conditions.

Contents

List of Abbreviations

ACP acyl carrier protein . . . 59
AICE actinomycete integrative and conjugative element . . . 56
ATP adenosine triphosphate . . . 75
BAC bacterial artificial chromosome . . . 11
cACPL cluster of *Actinoplanes* . . . 60
cDNA copy DNA . . . 16
CDD conserved domain database . . . 72
CDS coding sequence . . . 14
CDW cell dry weight . . . 63
COG cluster of orthologous groups of proteins . . . 25
DAPA *meso*-2,6-diaminopimelic acid . . . 1
DE differentially expressed . . . 64
DNA deoxyribonucleic acid . . . 2
EC enzyme commission . . . 25
emPCR emulsion PCR . . . 23
GC-MS gas-chromatography mass-spectrometry . . . 32
GOLD Genomes Online Database . . . 2
GS Genome Sequencer . . . 23
GUI graphical user interface . . . 15
HDAPA hydroxy-diaminopimelic acid . . . 1
HMM hidden Markov model . . . 26
HPA human pancreatic α-amylase . . . 6
HPLC high performance liquid chromatography . . . 30
KEGG Kyoto encyclopedia of genes and genomes . . . 25
MAT malonyl transferase . . . 62
mRNA messenger RNA . . . 16
ncRNA non-coding RNA . . . 17
NCBI National Center for Biotechnology Information . . . 2
NRPS non-ribosomal peptide synthetase . . . 59

ORF open reading frame . . . 14
OSMAC one strain, many compounds . . . 50
pACPL plasmid of *Actinoplanes* . . . 56
PCP peptidyl carrier protein . . . 59
PCR polymerase chain reaction . . . 12
PE paired-end . . . 12
PKS polyketide synthase . . . 59
rDNA ribosomal DNA . . . 15
RNA ribonucleic acid . . . 14
RNA-seq RNA-sequencing . . . 16
RNR ribonucleotide reductase . . . 74
RPKM reads per kilobase of coding sequence per million mapped reads . . . 31
rRNA ribosomal RNA . . . 14
TEN terminator exonuclease . . . 30
tmRNA transfer-messenger RNA . . . 72
TLS translation start . . . 66
TPP trehalose 6-phosphate phosphatase . . . 56
TPS trehalose 6-phosphate synthase . . . 56
tRNA transfer RNA . . . 14
TS trehalose synthase . . . 56
TSS transcription start site . . . 17
UTR untranslated region . . . 66
WGS whole genome shotgun . . . 11

List of Figures

List of Tables

1 Chapter 1. Introduction

1.1. The genus *Actinoplanes*

The genus *Actinoplanes* was first introduced by John Nathaniel Couch in 1950 with *Actinoplanes philippinensis* as its type strain [Couch, 1950]. Taxonomically, *Actinoplanes* is classified within the family Micromonosporaceae and order Actinomycetales, which belongs to the broad class of Actinobacteria. Species of that genus colonize various habitats including different soil, freshwater, and marine environments. They are distinguished from other members of the family Micromonosporaceae mainly through their characteristic formation of globose sporangia, containing globular spores, which become motile soon after dehiscence. Other distinctive characteristics are the usual absence of an aerial mycelium and the composition of the cell wall, which contains *meso*-2,6-diaminopimelic acid (DAPA), ll-2,6-diaminopimelic acid, and/or hydroxydiaminopimelic acid (HDAPA), and glycine [Lechevalier & Lechevalier, 1970]. Because of these components, the cell wall of *Actinoplanes* spp. belongs to the chemotype II and resembles that of Gram-positive bacteria [Šuput *et al.*, 1967]. Among different species the ratio of DAPA to HDAPA differs significantly, ranging from pure DAPA in *Actinoplanes philippinensis* through *Actinoplanes missourensis*, which has roughly equal amounts of both amino acids, to pure HDAPA in *Actinoplanes utahensis* [Parenti & Coronelli, 1979]. Another rare feature of the cell wall is the substitution of *N*-acetylmuramic acid by *N*-glycolylmuramic acid within the peptidoglycan layer, which explains the resistance of *Actinoplanes* strains against the *N*-acetylmuramide glycanhydrolase lysozyme [Vobis, 1989].

Most members of the genus *Actinoplanes* grow aerobically under mesophilic temperature conditions ranging from 15 to 37 °C with an optimum at around 30 °C. They feed saprophytically on dead plant material, pollen grains, and chitin-containing biological material which entails their good utilization of major components of plant cell walls such as xylose and arabinose. The majority of *Actinoplanes* species form dense colonies with regular shapes, whose central protuberance usually develops a straight sporangiophore supporting the characteristic sporangia. The colonies are typ-

ically of orange or yellow color, which is due to an aggregation of pigments within the protoplasm of the cells. These pigments were shown to be carotenoids that seem to require light for their synthesis. This is exceptional in the sense that non-photosynthetic bacteria and fungi are generally able to produce carotenoids irrespectively of light conditions [Parenti & Coronelli, 1979]. *Actinoplanes* spp. exhibit a genomic deoxyribonucleic acid (DNA) content of 70-73 mol% guanine-cytosine nucleotide bases, which is typical for Actinobacteria [Farina & Bradley, 1970]. This high *GC-content* has several implications on DNA-sequencing strategies, as described later.

Actinobacteria are a rich source for industrially and pharmacologically valuable compounds, such as antibiotics, amino acids, functional food additives, and drug precursors [Ventura *et al.*, 2007]. This is especially reflected by the rising number of genome sequencing projects dealing with members of this phylum. According to the Genomes Online Database (GOLD)[1] as of October 2011, about 11% of all sequencing projects work already with actinomycetes, which stresses the rising interest in secondary metabolites that are produced by their diverse species. The genus *Actinoplanes* fits well into this trend. It represents the richest group of the rare actinomycetes with at least 45 validly described species and more than 200 isolates listed in the taxonomy database of the National Center for Biotechnology Information (NCBI)[2] as of December 2011. In particular, more than 120 antibiotics have been reported from these species. Among these compounds, amino acid derivatives such as peptides and depsipeptides prevale [Lazzarini *et al.*, 2001]. Especially the glycopeptide teicoplanin produced by *Actinoplanes teichomyceticus* is of clinical relevance for the treatment of life-threatening infections by Gram-positive bacteria, particularly those caused by methicillin-resistant *Staphylococcus aureus* strains [Jung *et al.*, 2009]. Other antibiotics of elevated interest include lipiarmycin from *Actinoplanes deccanensis* [Parenti *et al.*, 1975], ramoplanin from *Actinoplanes* sp. ATCC 33076 [Cavalleri *et al.*, 1984], purpuromycin from *Actinoplanes ianthinogenes* [Kirillov *et al.*, 1997], and friulimicin from *Actinoplanes friuliensis* [Aretz *et al.*, 2000].

1.2. The strain *Actinoplanes* sp. SE50/110

On the 22^{nd} of December 1969, a new strain, designates *Actinoplanes* sp. SE50 (ATCC 31042; CBS 961.70), was isolated through pollen-baiting from a soil sample taken from a coffee plantation near the city of Ruiru in Kenya, Africa. Among other isolates, *Actinoplanes* sp. SE50 was tested in the course of a screening experiment for new substances with inhibitory effects on glycoside hydrolases by the company Bayer AG. The culture broth of the strain showed remarkable inhibitory effects on mammalian intestinal amylases, maltases, and saccharases and therefore became subject to further investigation [Frommer *et al.*, 1975]. In the following, it was found that the active compound of the broth was comprised of a mixture of complex oligosaccharides

[1] URL: http://www.genomesonline.org
[2] URL: http://www.ncbi.nlm.nih.gov/Taxonomy/

of which the pseudo tetrasaccharide *acarbose* was the most potent inhibitor of α-glucosidases [Schmidt *et al.*, 1977]. Later, a natural variant of the original wild-type isolate *Actinoplanes* sp. SE50, designated *Actinoplanes* sp. SE50/110 (ATCC 31044; CBS 674.73), was found to produce elevated levels of up to 1 g/L of acarbose [Frommer *et al.*, 1979]. Since then, *Actinoplanes* sp. SE50/110 (**Fig. 1.1**) has been used in many research studies which helped to identify and reveal the DNA sequence of the acarbose biosynthetic gene cluster as well as the functional characterization of its encoded enzymes [Crueger *et al.*, 1998a, Stratmann *et al.*, 1999, Hemker *et al.*, 2001, Zhang *et al.*, 2002, Zhang *et al.*, 2003].

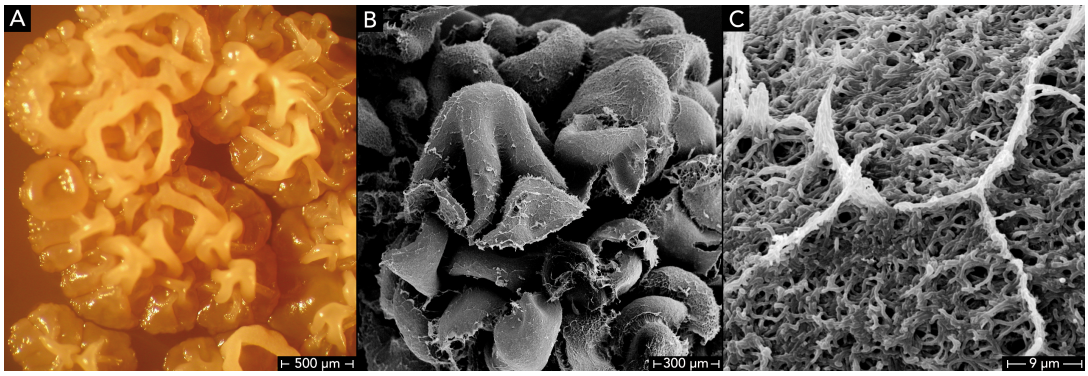

Figure 1.1.: Three images with different levels of magnification of an *Actinoplanes* sp. SE50/110 culture grown on agar plates. (**A**) Light microscopy image; (**B**) electron microscopy image with moderate magnification; (**C**) electron microscopy image with high magnification.

Of note, in a patent from the year 2000 the author introduced the species name *Actinoplanes acarbosefaciens* for all strains derived of *Actinoplanes* sp. SE50 [Crueger, 2000]. However, this name was not used in any scientific publication before and ever since.

1.3. The secondary metabolite acarbose, its relevance, and mode of action

The α-glucosidase inhibitor acarbose, *O*-{4,6-dideoxy-4[1s-(1,4,6/5)-4,5,6-trihydroxy-3-hydroxymethyl-2-cyclohexen-1-yl]-amino-α-D-glucopyranosyl}-(1$\rightarrow$ 4)-*O*-α-D-glucopyranosyl-(1$\rightarrow$4)-D-glucopyranose, is a special representative of a complex group of compounds, called amylostatins [Wehmeier & Piepersberg, 2004]. Its chemical structure is composed of a valienamine moiety which is *N*-glycosidically bound to 4-amino-4,6-dideoxyglucose, resulting in the *core* structure of the molecule, the pseudodisaccharide acarviosine (valienaminyl-4-amino-4,6-dideoxyglucose). Acarviosine is further α-1,4-linked to a maltose residue, constituting the complete acarbose unit (**Fig. 1.2**) [Müller *et al.*, 1980, Truscheit *et al.*, 1981].

Besides acarbose, *Actinoplanes* sp. SE50/110 produces a wide variety of other pseudooligosaccharides that all have the acarviosine core structure in common but

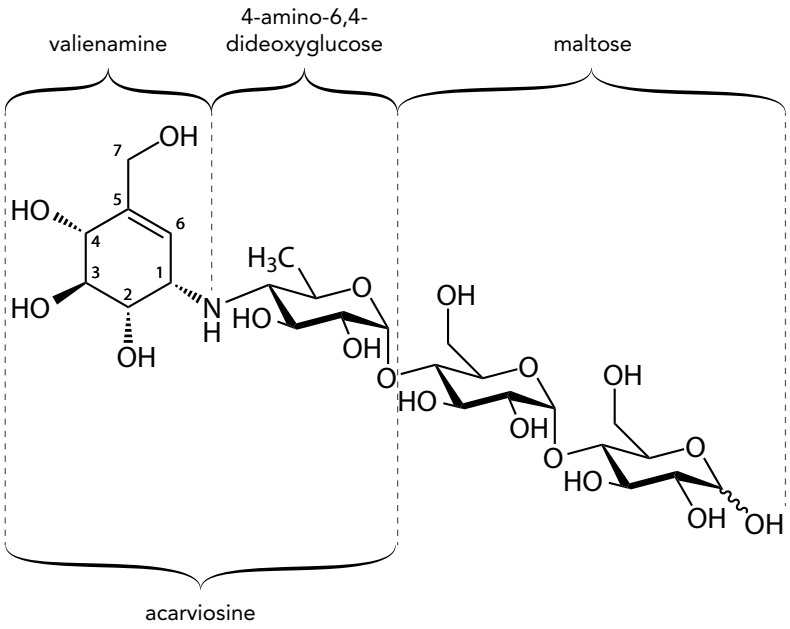

Figure 1.2.: The chemical structure of acarbose.

differ in the number, nature, and bond-type of the molecules at its reducing (R_n) and non-reducing (R_m) end (**Fig. 1.3**). While homologues of acarbose are characterized through the sole use of α-1,4-linked glucose molecules, other derivatives contain fructose, mannose, 1-*epi*,2-*epi*-valienol and vary in the terminal glycosidic bond [Müller *et al.*, 1980, Hemker *et al.*, 2001]. In this regard, component C (**Tab. 1.1**) is of special interest because of its structural similarity to acarbose, which renders the separation of both compounds challenging [Wehmeier & Piepersberg, 2004].

Table 1.1.: Names and compositions of acarviosyl-containing compounds

name	composition
Acarbose (component 3)	Acarviosyl-1-4-Glc-1-4-Glc
Component A	Acarviosyl-1-4-Glc-1-4-Fru
Component B	Acarviosyl-1-4-Glc-1-4-Val
Component C	Acarviosyl-1-4-Glc-1-1-Glc
Component D	Acarviosyl-1-4-Glc-1-4-Man
Component 4a	Acarviosyl-1-4-Glc-1-4-Glc-1-4-Fru
Component 4b	Acarviosyl-1-4-Glc-1-4-Glc-1-4-Glc
Component 4c	Acarviosyl-1-4-Glc-1-4-Glc-1-1-Glc
Pseudo-acarbose	Acarviosyl-1-4-(6-desoxy)Glc-1-4-Glc

The length of the oligosugars linked to the R_n and R_m ends are largely determined by the supplied carbon source in the cultivation broth and can vary between 1 and

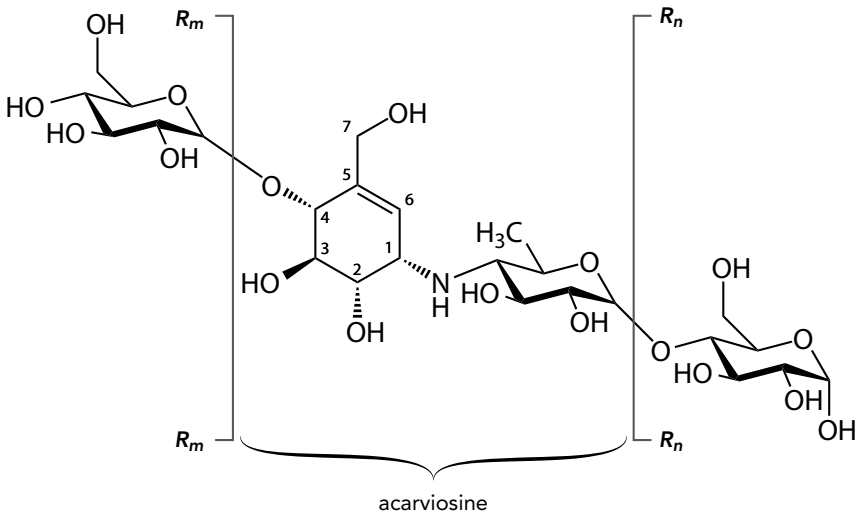

Figure 1.3.: Chemical structure of acarbose homologues.

30 units. While acarbose and other shorter pseudooligosaccharides are preferably produced in maltose and glucose containing media, already small amounts of supplied starch lead to the production of longer products [Schmidt *et al.*, 1977, Frommer *et al.*, 1979]. This is crucial, as the number of glucose molecules bound to acarviosine influences the substrate specificity of the compound as an inhibitor. Low-molecular compounds such as acarbose and component 2 possess strong inhibitory effects on disaccharases, whereas high-molecular compounds are more effective in inhibiting α-amylases [Frommer *et al.*, 1979, Müller *et al.*, 1980].

The inhibitory effect of all acarbose-related pseudooligosaccharides is based on their inherent acarviosine core structure. In contrast to α-1,4-glycosidic bonds, the *N*-glycosidic linkage between valienamine and the 4-amino-4,6-dideoxyglucose can not be hydrolyzed by the catalytic centers of α-glucosidases [Heiker *et al.*, 1981]. Rather, they simulate an intermediate state in the cleavage process of these enzymes which is why they are also known as *transition-state-analogues* [Habermehl *et al.*, 2008]. X-ray studies on a sucrase-isomaltase complex, isolated from the small intestine of rats, first revealed the competetive mechanism of the inhibition [Sigrist *et al.*, 1975, Hanozet *et al.*, 1981, Samulitis *et al.*, 1987]. Further kinetic studies indicated that the sucrase possessed a 15,000-fold higher affinity to acarbose in comparison to its natural substrate sucrose, which ultimately leads to the inhibition of the enzyme [Caspary & Graf, 1979]. The inhibitory effect of acarbose on human α-glucosidases of the small intestine was discovered by Caspary and coworkers who also noted its potential application in the treatment of type-2 diabetes mellitus [Caspary & Graf, 1979, Caspary & Kalisch, 1979].

Diabetes mellitus type-2 is a chronic disease with more than 250 million people affected worldwide. Inappropriately managed or untreated, it can lead to severe cases of renal failure, blindness, slowly healing wounds, and arterial diseases, including coro-

nary artery atherosclerosis [IDF, 2009]. The underlying cause for this disease is a concurrent deficit of insulin secretion or insulin action and insulin resistance, which results in hyperglycemia due to the reduced ability to absorb and use glucose in the muscles and in the liver [Bottino & Trucco, 2005]. Acarbose specifically aids in the development and control of hyperglycemia by reducing the uptake rate of glucose in the human intestinal tract. This is achieved by the aforementioned inhibition of α-glucosidase enzymes in the brush border of the small intestine, which, in the absence of acarbose, would rapidly degrade oligosaccharides, trisaccharides, and disaccharides into glucose and other monosaccharides whose massive absorption leads to pathogenic blood sugar levels. Another important effect of acarbose is its inhibition of the human pancreatic α-amylase (HPA) in the lumen of the small intestine, which reduces the rate by which complex starches are hydrolized to oligosaccharides [Wehmeier & Piepersberg, 2004]. Several α-amylases including HPA also possess the ability to convert acarbose into even more efficient inhibitors through different transglycosylation reactions between two or more acarbose molecules. This results in a refinement of the molecular positioning of the N-glycosidic bond of acarviosine within the active center of the α-glucosidases and increases their affinity to the inhibitor [Dauter *et al.*, 1999, Nahoum *et al.*, 2000]. Thus, acarbose acts at least in some of its target enzymes as a prodrug [Wehmeier & Piepersberg, 2004].

In order to reduce the postprandial hyperglycemia in diabetes patients after ingestion of carbohydrate-containing diets by 50%, a dose of 1-1.5 mg of acarbose per kg body weight is advisable [Truscheit *et al.*, 1988]. The usual dosage form is in white tablets that are taken with the first bite of food intake to develop the optimal inhibitory effect.

Besides its application in diabetes mellitus type-2, acarbose was also tested for its applicability in other medical fields such as obesity, adipose, hyperlipidemia (arteriosclerosis), gastritis, gastric ulcer, duodenal ulcer, and caries in man, or as food additive for various purposes in farm animals [Frommer *et al.*, 1977a, Frommer *et al.*, 1977b, Frommer *et al.*, 1979, Frommer *et al.*, 1975].

1.4. The biosynthesis of acarbose in *Actinoplanes* sp. SE50/110

The foundation for the genetic analysis of the acarbose biosynthesis was laid in 1992, when DNA probes were designed from the streptomycin biosynthesis genes *strDELM* of *Streptomyces griseus*. Like many other secondary metabolites of Actinobacteria, the biosynthesis of streptomycin requires enzymes responsible for the synthesis of precursor compounds via the highly conserved dTDP-hexose pathway. It was therefore possible to use these DNA probes in screening experiments for homologous genes in related bacteria [Stockmann & Piepersberg, 1992].

The same approach was successfully applied for *Actinoplanes* sp. SE50/110 using the DNA probes for the dTDP-glucose 4,6-dehydratase gene *strE* and resulted in the identification of the gene *acbB*, which likewise encodes a dTDP-glucose 4,6-dehydratase [Crueger *et al.*, 1998b]. For the reason that secondary metabolite genes were often found to be organized in clusters, further work concentrated on the

cloning and sequencing of adjacent genomic DNA regions of this gene [Stratmann *et al.*, 1999, Hemker *et al.*, 2001, Wehmeier, 2003]. These efforts lead to the identification of the acarbose biosynthesis gene cluster [GenBank:Y18523.4] (**Fig. 1.4**). The 32 kb long *acb* gene cluster consists of 22 genes which are organized in at least eight transcriptional units, namely *acbZ*, *acbWXY*, *acbVUSRPI*, *JQKMLNOC*, *acbB*, *acbA*, *acbE*, and *acbD* [Thomas, 2001]. According to latest findings, the operon *acbHFG* is not directly involved in the biosynthesis or the metabolism of acarbose [Licht *et al.*, 2011], which is in contrast to earlier assumptions [Brunkhorst *et al.*, 2005].

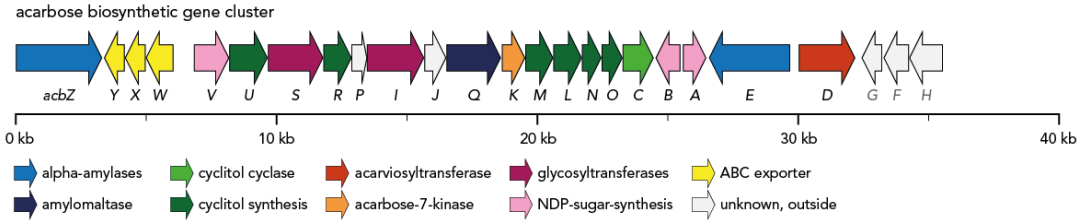

Figure 1.4.: The acarbose biosynthetic gene cluster of *Actinoplanes* sp. SE50/110.

The functions of most gene products have already been elucidated, which gives a fairly complete picture of the biosynthesis and metabolism of acarbose in *Actinoplanes* sp. SE50/110 (**Fig. 1.5**). As a first step in the synthesis of the valienamine subunit of acarbose (**Fig. 1.2**), *sedo*-heptulose-7-phosphate, which originates from the penthose phosphate pathway, is cycled by the C7-cyclitol synthase AcbC to form 2-*epi*-5-*epi*-valiolone [Stratmann *et al.*, 1999]. This intermediate is subsequently C7-phosphorylated by the ATP-dependent kinase AcbM, yielding 2-*epi*-5-*epi*-valiolone-phosphate [Zhang *et al.*, 2002]. Notably, this phosphorylation is maintained throughout the complete biosynthesis of acarbose and protects the own cytosolic enzymes against inhibition by acarbose [Drepper & Pape, 1996, Goeke *et al.*, 1996]. Next, AcbO catalyzes the C2-epimerization to 5-*epi*-valiolone-7-phosphate [Zhang *et al.*, 2003], which is then reduced to 5-*epi*-valiolol-7-phosphate by the NADH-dependent dehydrogenase AcbL [Wehmeier, 2003]. The following dehydratase reaction to 1-*epi*-valienol-7-phosphate is driven by AcbN, which belongs to the family of short-chain oxidoreductases. For the next step, the C1-phosphorylation to 1,7-diphospho-1-*epi*-valienol, a responsible enzyme has not been determined yet. However, AcbU – a putative cyclitol kinase – is a likely candidate [Wehmeier & Piepersberg, 2004]. The subsequent nucleotidylation to NDP-1-*epi*-valienol-7-phosphate is presumably catalyzed by the ADP-glucose synthase AcbR [Wehmeier, 2003].

In contrast to the synthesis of the valienamine precursor, all catalytic steps of the deoxysugar moiety of acarbose are well characterized and follow the dTDP-hexose pathway mentioned above [Liu & Thorson, 1994, Piepersberg & Distler, 1997]. It starts with the nucleotidylation of D-glucose-1-phosphate to dTDP-D-glucose by the dTDP-glucose synthase AcbA and is then further modified by the above mentioned dTDP-glucose 4,6-dehydratase AcbB, which results in the formation of dTDP-4-keto-6-deoxy-D-glucose [Wehmeier, 2003]. The next step yielding dTDP-4-amino-

4,6-dideoxy-D-glucose is catalyzed by the aminotransferase AcbV, which utilizes L-glutamic acid as donor for the amino group [Piepersberg *et al.*, 2002].

The concatenation of the subunits NDP-1-*epi*-valienol-7-phosphate and dTDP-4-amino-4,6-dideoxy-D-glucose to dTDP-acarviose-7-phosphate is finally catalyzed by the glycosyltransferase-like protein AcbS [Wehmeier & Piepersberg, 2004]. Earlier feeding experiments showed that the following addition of a maltose moiety originates from free maltose or maltotriose rather than from successive addition of two glucose molecules which were only utilized for the formation of dTDP-4-amino-4,6-dideoxy-D-glucose [Lee *et al.*, 1997]. The enzyme responsible for the condensation of maltose and dTDP-acarviose-7-phosphate to acarbose-7-phosphate has not yet been determined, however. A likely candidate that could fulfill this function is the second glycosyltransferase-like protein AcbI [Wehmeier & Piepersberg, 2004, Rockser & Wehmeier, 2009]. Acarbose-7-phosphate is then exported into the extracellular medium, presumably through the ABC-transporter AcbWXY, which is also held responsible for the dephosphorylation and, hence, the activation of acarbose [Piepersberg *et al.*, 2002].

In its natural environment, acarbose is believed to play a multifunctional role in the acquisition of glucose-containing carbon sources for *Actinoplanes* sp. SE50/110. First, it inhibits starch degrading enzymes of nutrient competitors and their maltodextrine uptake systems. Second, it serves as an acceptor of oligosugars, which are provided through starch degradation by the own secreted acarbose-resistant α-amylases AcbE and AcbZ and subsequent transfer to acarbose by the acarviosyltransferase AcbD. Third, it serves as a recyclable transport vehicle for these loaded acarbose compounds, which are eventually imported through a yet to be determined importer complex. Previously it was assumed that AcbHFG takes this role [Brunkhorst *et al.*, 2005], but recent findings exclude that possibility as it was demonstrated that the extracellular binding protein AcbH possesses a predominant specificity for galactose [Licht *et al.*, 2011]. The loaded acarbose is intracellularly recycled through deglycolization by the amylomaltase AcbQ and thus, the cleaved glucose molecules are available for utilization in the central metabolism of *Actinoplanes* sp. SE50/110. Immediately afterwards, acarbose is re-phosphorylated by the acarbose-7-kinase AcbK, which serves the same purpose as during its synthesis – the protection of own intracellular enzymes and tagging for re-export through AcbWXY.

This proposed *carbophore*-cycle elegantly explains how acarbose (and presumably its homologues) facilitates the life of its producer in a community which competes for the same carbon sources [Wehmeier & Piepersberg, 2004]. Another putative function of the carbophore could be related to *quorum sensing* (reviewed in [Schauder & Bassler, 2001]) in that it would enable *Actinoplanes* sp. SE50/110 to measure the population density in its environment [Piepersberg, 1993, Piepersberg *et al.*, 2002].

The exact functions of the proteins AcbP and AcbJ are currently unknown, although AcbJ might be involved in the dephosphorylation of acarbose-7-phosphate during the export through AcbWXY [Wehmeier & Piepersberg, 2004]. For AcbP, which shows some sequence similarity to the NUDIX-hydrolase family, a putative reg-

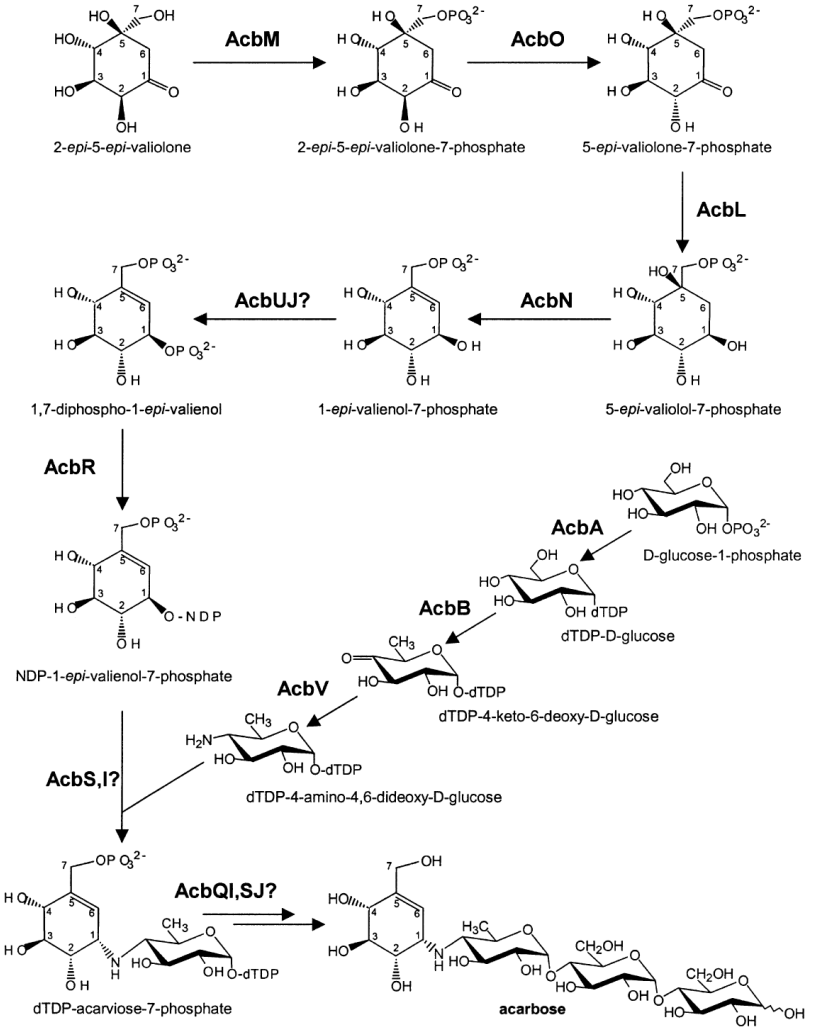

Figure 1.5.: Postulated pathways of the acarbose biosynthesis in *Actinoplanes* sp. SE50/110 [Zhang *et al.*, 2002].

ulatory function on the metabolic level has been proposed. This regulation would e.g. involve the hydrolization of accumulated toxic NDP-1-*epi*-valienol-7-phosphate to 1-*epi*-valienol-7-phosphate which might subsequently be dephosphorylated by AcbJ to form 1-*epi*-valienol. In this theory, 1-*epi*-valienol accumulates as a result of this detoxification as previously observed [Mahmud *et al.*, 1999].

Interestingly, a second acarbose biosynthetic gene cluster *gac* was recently identified and characterized in *Streptomyces glaucescens* GLA.O (DSM 40716). It exhibits remarkable similarities to the *acb*-cluster but differs in the synthesis of 1-*epi*-valienol-7-phosphate after the initial cyclization step from *sedo*-heptulose-7-phosphate to 2-*epi*-5-*epi*-valiolone. More importantly, the cluster contains two putative transcriptional regulators, GarC1 and GarC2, for which no homologues exist in the *acb*-cluster. One

or both of these regulators are assumed to control the expression of the AcbHFG homologue ABC-importer GacHFG. In contrast to AcbH, GacH might be able to import loaded acarbose in the sense of the carbophore model, as its substrate specificity is likely to be different [Rockser & Wehmeier, 2009]. However, this has not yet been demonstrated.

Very few information exists on the regulation of the *acb* gene cluster. Yet it was found that acarbose production in *Actinoplanes* sp. SE50/110 starts during the exponential growth phase and not, like most other antibiotics, in the stationary phase. This suggests a regulation coupled with the carbohydrate metabolism as the need for energy is highly increased during the exponential growth phase [Drepper & Pape, 1996, Piepersberg & Distler, 1997]. Other experiments indicated a substrate-induced regulation of AcbE and AcbD after addition of maltose and maltotriose to the culture medium [Stratmann, 1997], which is in line with experiments that found the complete *acb*-cluster induced after addition of maltotriose and higher maltooligosaccharides (up to maltoheptaose) [Wehmeier, 2003].

1.5. Industrial development and fermentation of acarbose

Since its initial authorization in Switzerland in the year 1986, acarbose emerged as a widely used drug in the treatment of diabetes mellitus type-2 and now belongs to the ten top-selling pharmaceuticals produced by the company Bayer HealthCare AG. In 2010, its annual sales volume raised to 347 million Euro and an increasing demand, especially from Asian countries, is anticipated. Today, acarbose is sold in 110 countries worldwide under different tradenames, such as Glucobay®(Europe and China), Precose®(United States), Glucor®(France), and Prandase®(Canada) [Bayer AG, 2011].

The industrial production of acarbose is established as a multi-step fed-batch fermentation with strains derived from *Actinoplanes* sp. SE50. In order to increase cost-efficiency and compete with raising demands over the years, laborious conventional mutagenesis and screening experiments were conducted with the aim to develop strains with increased acarbose yields. This long lasting optimization procedure was very successful in that the latest production strains produce ∼500-fold more acarbose than the original isolate [Schedel, 2006]. Concurrently, the development of the production media was pushed forward, whose composition and adaptation to new production mutants can have significant impact on acarbose production efficacy [Schmidt *et al.*, 1977, Frommer *et al.*, 1979].

Downstream processing of the acarbose fermentation broth is accomplished by a multi-stage purification process. It includes a series of highly specialized chromatographic columns and enrichment steps before the dried acarbose powder reaches its final purity of >98% [Wehmeier & Piepersberg, 2004, Schedel, 2006].

A major challenge in the purification of acarbose is its separation from the structurally similar derivative component C (**Tab. 1.1**). This compound also accounts for most of the remaining <2% impurity of the product. In total, the complete biotechnological process requires two to three weeks [Wehmeier & Piepersberg, 2004].

1.6. Bacterial genome sequencing approaches

Since the discovery of the double helical structure of the DNA molecule on April 15, 1953 by James Watson and Francis Crick, researchers around the world have strived to develop methods that allow the reading of the genetic code out of a DNA molecule [WATSON & CRICK, 1953]. Early successful methods include plus-minus sequencing of Frederick Sanger and Alan Coulson [Sanger & Coulson, 1975] as well as the method of Allan Maxam and Walter Gilbert, which was based on chemical modification and subsequent base specific cleavage [Maxam & Gilbert, 1977]. Because of the extensive use of hazardous chemicals and since the introduction of the improved chain-termination method in Sanger sequencing, the Maxam-Gilbert method soon disappeared. Several further improvements, such as fluorescently labeled ddNTPs, the introduction of dye-terminators, capillary electrophoresis, and automatization lead to a widespread application of this method. While sequencing of DNA fragments soon became a standard method in many laboratories, the sequencing of an entire bacterial genome still posed a challenging, expensive and time consuming task. The hierarchical standard procedure involved the clonal amplification and storage of larger DNA pieces of the target genome in bacterial artificial chromosome (BAC) libraries. These were screened for a minimal set that contained DNA fragments that together constituted the complete target genome (minimal tiling path). The actual sequencing was performed using chromosome walking, which progresses through the inserted DNA fragment of a BAC in a sequential order and yields one *read* of approximately 600-1000 bases in length at a time [Chinault & Carbon, 1979, Niedringhaus *et al.*, 2011].

Definition 1 (read) *A* read *is a single contiguous sequence of nucleotide letters that are read from a DNA molecule.*

With decreasing sequencing costs, an alternative method, termed whole genome shotgun (WGS) sequencing, was applied. In shotgun sequencing, the target DNA is randomly fragmented into shorter pieces that undergo Sanger sequencing. The advantages in ease of use however, came at the cost of a necessary assembly step in which all overlapping reads are aligned in order to yield a *contig* [Staden, 1979].

Definition 2 (contig) *A* contig *is a contiguous consensus region of DNA that is inferred from a set of overlapping reads.*

Even today, enhanced devices with up to 384 parallel reactions are still based on Sanger's chain-termination principle and are frequently used for specialized applications that do not require high-throughput yields [Hert *et al.*, 2008].

A new era of DNA-sequencing began in 2005 with the first commercially available second generation device, the *Genome Sequencer 20* marketed by 454 Life Sciences [Margulies *et al.*, 2005]. In contrast to the first generation of Sanger sequencing, second generation refers to a type of sequencing that does not require BAC cloning

for *de novo* sequencing but runs automatically based on enzymological amplification in a massively parallel manner [Mardis, 2008b].

Without doubt, the introduction of second generation sequencing devices has essentially contributed to the wealth of today's available genome sequences from all domains of life. Featuring long read length, unprecedented low costs per sequenced base, and high-throughput in short time, *pyrosequencing* (454 Life Sciences, a Roche company) and *sequencing-by-synthesis* (Illumina) are currently the most commonly applied high-throughput sequencing technologies [Ahn, 2011]. Although both methods are well suited for the elucidation of complete genome sequences, they bear decisive differences. While the Ilumina platform yields considerably more sequence information per run, the resulting reads are of short length compared to those usually obtained from Roche/454 devices. Read length is an important parameter, especially in *de novo* genome sequencing because it determines the maximal length of repetitive genomic elements that can be unambiguously determined by the technology. Depending on the genome under investigation, the difference in read length between 100 bp (Illumina) and 500 bp (Roche/454) can have severe impact on the following assembly process, in that the Illumina technology might result in much more contigs than the assembly of the Roche/454 data [Mardis, 2008a]. On the other hand, Illumina's superior throughput is usually beneficial for re-sequencings of genomes with a known or closely related reference sequence as it yields higher coverage and the costs per base are considerably less expensive.

While hundreds of genomes have been successfully sequenced with these technologies, most did not result in the reconstruction of the complete genome due to gaps between contiguous sequences [Eichler *et al.*, 2004, Chain *et al.*, 2009]. To finish these genomes nonetheless, several time- and cost-intensive tasks have to be carried out. These include the arrangement of contigs into the order of their natural occurrence, for which clone libraries have to be constructed and end-sequenced, covering the whole genome with overlapping stretches of genomic DNA. Subsequent primer picking, polymerase chain reaction (PCR) amplification, and Sanger (capillary) DNA-sequencing of the products finally result in the determination of the gap sequences, which have to be manually added to the assembly afterwards in order to completely finish the genome [Tauch *et al.*, 2008, Trost *et al.*, 2010].

These finishing steps can drastically increase the project costs and duration while the amount of new DNA sequence information is only marginal. It is therefore comprehensible that an increasing number of genome projects omit the finishing process, leaving the genome in draft status [Chain *et al.*, 2009]. However, many advanced and in particular, comparative genome analyses can hardly be applied to such heavily fragmented datasets [Fraser *et al.*, 2002]. For these reasons it is desirable to find new ways of improving the automated sequencing pipelines in order to yield higher quality genomes with less effort. Several of these improvements, such as optical mapping [Latreille *et al.*, 2007], increased read length, and paired-end (PE) sequencing have already proven to enhance the genome quality considerably as well as allowing for new analytic methods [Bashir *et al.*, 2008].

Both Illumina and Roche/454 technologies provide the opportunity to perform PE sequencing. In PE sequencing, both ends of the same larger DNA fragment are sequenced. The advantage of this method is that the distance between the two sequenced ends is approximately known, which allows the assembly software to build a *scaffold* of the genome in which the order and orientation of the contigs are known.

Definition 3 (scaffold) *A* scaffold *is a set of non-overlapping contigs in which the order and orientation of all contigs are known.*

The gaps between the contigs in a scaffold can then be closed by the addition of reads from e.g. a second ordinary sequencing run. During the finishing phase, the last remaining gaps are usually closed by manual assembly of reads obtained from Sanger sequencing of genomic PCR products. Sometimes complex genomes may also necessitate the construction of a fosmid library which aids in the finishing phase of the project in that it allows the PCR-based amplification and subsequent sequencing of isolated DNA regions. In particular, larger repetitive regions, such as ribosomal operons, can be individually sequenced using fosmid libraries. Moreover, it has been generally observed that second generation sequencing of high GC-content material tends to be much more difficult as it results in significantly more gaps than average or low GC-content sequences [Frey *et al.*, 2008]. In these cases, fosmid libraries can be used to amplify and sequence uncovered regions with specialized PCR protocols [Sahdev *et al.*, 2007] and reagents [Turner & Jenkins, 1995, Spiess *et al.*, 2004, Henke *et al.*, 1997].

Another aspect of second generation sequencing is the vast amount of data that can be generated by these devices. For instance, a single run on one HiSeq 2000 device (Illumina) yields between 540 and 600 gigabases of sequence information [Ahn, 2011], which allows for two human genomes to be sequenced with a 30-fold coverage in a single run. Storing and processing these data poses an increasingly challenging task, which drives the constant development of new and highly efficient data formats and algorithms. Especially in modern assembly software a clear shift is observed, away from using overlap-graphs to tools employing more memory efficient De Bruijn graphs as implemented in `ABySS` [Simpson *et al.*, 2009], `SOAPdenovo` [Li *et al.*, 2008], and `Velvet` [Zerbino & Birney, 2008].

Besides the Roche/454 and Illumina sequencing technologies discussed above, other noteworthy platforms are briefly mentioned here. The other two second generation technologies are ABI/SOLiD's *sequencing by ligation* method that yields very high throughput at read lengths of 25-35 bases, and HeliScope's *single molecule sequencing by synthesis* which offers high throughput and read lengths of 25-30 bases. Also third generation platforms have recently reached the market. These include the *real-time single-molecule* sequencer PacBio RS (Pacific Biosciences) and Ion Torrent's *Personal Genome Machine*, which uses semiconductor sequencing technology. Furthermore, Oxford Nanopore's gridION platform is considered as the first technology of the fourth generation [Niedringhaus *et al.*, 2011].

1.7. Bacterial genome annotation strategies

Upon successful completion of a sequencing and assembly phase of a genome, an annotation procedure follows. The foremost task is usually the identification of protein coding sequences (CDSs) by specialized gene prediction software. These *gene finders* can be categorized into three groups. First, intrinsic gene finders rely only on previously incorporated rules and the genome sequence itself to predict CDS regions. These rules may include, as an example, a list of sequence motifs with known functions such as ribosomal binding sites and start/stop codons, or may apply statistical methods to derive coding probabilities based on local sequence composition [Ishikawa & Hotta, 1999]. Well-known representatives include `Glimmer` [Salzberg *et al.*, 1998,Delcher *et al.*, 1999], `GeneMark` [Besemer *et al.*, 2001, Besemer & Borodovsky, 1999], and `Prodigal` [Hyatt *et al.*, 2010]. Second, extrinsic gene finders rely on evidence from external sources such as sequence databases. They apply sequence comparison methods that infer the coding probability of open reading frames (ORFs) from their homology to database sequences. One added benefit of extrinsic gene finders is that they can immediately assign a putative function to CDSs based on the stored annotations of the most similar genes from the database. Third, hybrid gene finders combine the benefits of both approaches in order to deliver improved results. Examples for these are `Critica` [Badger & Olsen, 1999] and `Orpheus` [Frishman *et al.*, 1998]. Most of today's gene prediction tools follow this idea in which extrinsic methods are applied first and sequence regions without annotated genes are scanned afterwards with intrinsic methods [Mathé *et al.*, 2002].

Besides gene detection, a genome annotation comprises many other analysis methods. For instance, the prediction of transmembrane helices in CDS regions can provide information about the encoded enzymes being membrane bound or constitute transmembrane proteins, such as porins or surface receptors [Sonnhammer *et al.*, 1998, Krogh *et al.*, 2001]. By this means, the cellular localization of a protein can be estimated, which is especially interesting for secreted proteins which usually possess N-terminal signal peptides that can be recognized by tools such as `SignalP` [Nielsen *et al.*, 1997, Nielsen & Krogh, 1998, Bendtsen *et al.*, 2004]. Other widely used prediction methods perform the identification of ribonucleic acid (RNA) genes, such as transfer RNAs (tRNAs) [Lowe & Eddy, 1997] and ribosomal RNAs (rRNAs) [Lagesen *et al.*, 2007]. Moreover, the prediction of Roh-independent transcription terminators by `transTermHP` aids in deciphering operon structures [Kingsford *et al.*, 2007a]. Functional annotation is added by sequence comparison to a variety of data repositories, such as protein databases [Apweiler *et al.*, 2004], the conserved domain database [Marchler-Bauer *et al.*, 2011a], and the database of protein families [Finn *et al.*, 2009]. Moreover, it is possible to determine the genomic location of the origin of replication [Gao & Zhang, 2008] and to predict putative operon structures [Salgado *et al.*, 2000].

For convenience and ease of administration, most of these predictive tools are incorporated into sophisticated *annotation pipelines* which successively execute the individual programs automatically in a highly parallelized way on compute clusters. Some

popular annotation pipelines are GenDB [Meyer *et al.*, 2003], MAGPIE [Gaasterland & Sensen, 1996], SABIA [Almeida *et al.*, 2004], and GeneVar [Yu *et al.*, 2007]. In addition, there is an increasing trend towards outsourcing genome annotations to large sequencing centers and database providers which is caused by the increasing costs for bioinformatic infrastructure and staff required to maintain and update these systems [Cantarel *et al.*, 2008].

Based on the results of the automated annotation procedures, more specialized analysis tools exist. Among them, comparative genomics methods are used to elucidate the differences and similarities of two or more annotated genomes on a higher level of abstraction. The comparative genomics suite EDGAR for example, allows the determination of the *core genome* from a set of given genomes [Blom *et al.*, 2009]. The core genome includes the genes that were observed in all inspected genomes with a high sequence similarity to each other and thus, are likely to encode proteins involved in essential cellular functions. Moreover, it is possible to calculate the *pan genome* (a set of all genes from all genomes without replicates) and *singletons*, i.e. genes that occur in only one of the genomes.

The construction of phylogenetic trees, which represent the evolutionary relationships between the genomes under investigation, can also be derived from comparative genomics data. This task usually involves the construction of a multiple sequence alignment by tools such as ClustalW [Thompson *et al.*, 2002] or MUSCLE [Edgar, 2004b, Edgar, 2004a] and subsequent inference of the evolutionary distances by the PHYLIP package [Retief, 2000], MEGA [Tamura *et al.*, 2007], or related software. The most widely used application is the visualization of multiple sequence alignments from 16S ribosomal DNA (rDNA) gene sequences. Based thereon, attempts have been made to visualize the tree of life, which is currently consisting of about 1 million 16S rDNA sequences [Cole *et al.*, 2009].

Oftentimes, the most accurate way of annotating genes and other genomic features is the manual inspection of regions of interest. Therefore, software tools were developed that feature a graphical user interface (GUI), which eases the manipulation of automatically assigned annotations. The most prominent example is certainly the program Artemis [Rutherford *et al.*, 2000], which is actively developed by the Sanger Institute. But also some annotation pipelines such as GenDB, support manual annotation via the web interface, which enables collaborative efforts in optimizing an annotation [Meyer *et al.*, 2003].

1.8. Means of bacterial transcriptome analysis

The transcriptome of a cell is the set of all RNA molecules that are present at a specific timepoint. This implies that the transcriptome, in contrast to the genome, is constantly changing during the life of an organism. Alterations to the transcriptome can be triggered by various causes that are often associated with modifications in the environment of the cell. Examples are changing temperatures, oxygen availability, humidity, or lighting conditions. More intrinsic reasons may include the growth phase

of the cell and the state of its life-cycle. Thus, the timepoint of RNA extraction from the cell is critical in all transcriptome experiments.

Another noteworthy consideration regarding transcriptome analysis is the turnover time of RNA molecules. It is known that ribosomal and transfer RNAs are comparatively stable, whereas the half-lifes of messenger RNAs is rather short, which leads to rapid degradation by ribonuclease enzymes [Deutscher, 2006]. However, multiple cellular mechanisms have been discovered that can increase the half-lifes of specific messenger RNAs (mRNAs) [Kushner, 2002]. For these reasons, it is important to perform an RNA isolation as quickly as possible to avoid the degradation of short-lived transcripts and consequently, the over representation of long-living ones. Moreover, as the total amount of mRNA is certainly less than 5%, of the total RNA in a cell, long isolation times will favor stable rRNAs and tRNAs to be yielded [Deutscher, 2003].

What is more, RNA is usually not isolated from a single cell but from a population whose individuals would certainly differ to quite some extend in their transcriptomes, if isolated from their natural habitat. These fluctuations are minimized however, if the population is grown under controlled conditions as is usual in modern laboratories. In particular, it is important to ensure that all cells of a culture have the same access to nutritions and oxygen (if aerobic), which is generally done by shaking or stirring the culture. However, mycel-forming bacteria like *Actinoplanes* sp. SE50/110 are different in that their populations do not grow as individual cells but are highly interconnected and tend to form tiny globules in shaking cultures. It has to be assumed that oxygen and nutrient supply is not equally distributed and hence the transcriptomes may vary.

Depending on the intention of the experiment, different analysis strategies can be applied. Comparative analysis of different cultivation media, for example, allow the measurement of transcriptional responses to the different media compositions. This enables the identification of individual regulated genes or gene clusters for each condition tested. While these experimental setups only give a single snapshot of the cultivation, often time series analyses are performed, where samples are taken at constant timepoints throughout a cultivation process. Consequently, these data allow to follow the expression of a gene or gene cluster over time, which might help to better understand the dynamics of gene regulation.

Until recently, microarray technology was the uncontested choice for performing whole-transcriptome experiments. However, in the advent of second generation sequencing technologies (see **Section 1.6**), a novel method, termed *RNA-sequencing*, emerged that bears many new opportunities for transcriptional studies. Previous methods, such as Northern blotting, quantitative reverse transcription polymerase chain reaction (qRT-PCR), and microarrays relied on the hybridization principle. More precisely, they measured continuous intensities of targeted oligonucleotides hybridizing to a particular locus for their sequence specificity. RNA-sequencing (RNA-seq) on the other hand, introduces the measure of discrete read counts whose specificity is given by sequence alignment matches. In detail, isolated RNA is converted into copy DNA (cDNA) by reverse transcription. Then the cDNA is sequenced with high-throughput technologies as described above and the resulting reads are aligned either

to a reference genome or assembled *de novo*. The discrete amount of reads that overlap a CDS after the alignment phase is considered to be the expression value of that gene.

In principle, RNA-seq can be used with any high-throughput sequencing technology and the Illumina [Nagalakshmi *et al.*, 2008], Roche/454 [Vera *et al.*, 2008], and ABI/SOLiD [Cloonan *et al.*, 2008] systems have already been applied for this purpose. It is noteworthy however, that read length is not as beneficial in RNA-seq as it is in genome sequencing. It is rather desired to obtain a high coverage (also known as *sequencing depth*) to increase the reliability of detecting rare, yet physiologically relevant RNA species [Mortazavi *et al.*, 2008, Croucher & Thomson, 2010]. Therefore the Roche/454 technology might not be the best choice for RNA-seq experiments.

One of the main advantages of RNA-seq over micoarrays is its ability to detect transcripts independent of the availability of the underlying genome sequence. Because the acquired reads can be assembled to CDSs, which in turn can be annotated to analyze their putative functions, an expensive sequencing of the genome may be avoided [Garber *et al.*, 2011]. Furthermore, RNA-seq is able to detect transcription boundaries at a single nucleotide resolution, as well as single nucleotide polymorphisms (SNPs) and other sequence variations. In contrast to microarrays, RNA-seq does not have any significant background noise, nor can its signals be over-saturated [Wang *et al.*, 2009, Costa *et al.*, 2010].

Based on these intrinsic advantages, RNA-seq offers an array of new applications beyond the measurement of gene expression levels. These include the ability to detect novel transcripts, the mapping of transcription start sites (TSSs), non-coding RNA (ncRNA) profiling, and the possibility to perform strand-specific RNA-seq for the detection of antisense transcripts [Ozsolak & Milos, 2011].

As with every new technology, some controversy exists also with regards to RNA-seq. It has been reported that the sample preparation tends to introduce biases which might inflect biological data. In particular, the involved reverse transcription, the sample fragmentation and the PCR amplification steps are considered to be prone to introducing variations [Bullard *et al.*, 2010, Wang *et al.*, 2009]. Likewise, the GC-content of transcripts was shown to bias the detection of transcripts in Illumina data, with high-GC transcripts being overrepresented in the resultset [Dohm *et al.*, 2008]. On the other hand, many studies demonstrated high reproducibility of experiments with this technology [Bainbridge *et al.*, 2006, Mortazavi *et al.*, 2008, Hashimoto *et al.*, 2009]. Moreover, constant inventions, such as the amplification-free Helicos single molecule sequencing [Thompson & Milos, 2011], and the development of direct RNA-seq methods without the detour over cDNAs will likely diminish current concerns in the future [Ozsolak *et al.*, 2009].

Two typical workflows for RNA-seq data generation and analysis are depicted in **Figure 1.6**. The first workflow shows the main steps necessary to improve the underlying genome annotation (**Fig. 1.6A**). A 5'-enriched cDNA library is advantageous for the later TSS detection. After the reads were obtained by RNA-seq, they are either aligned to a reference genome by efficient mapping software, such as `Bowtie` [Lang-

MEAD *et al.*, 2009], `BWA` [LI & DURBIN, 2009], and `SARUMAN` [BLOM *et al.*, 2011] or assembled *de novo* with programs, such as `Cufflinks` [TRAPNELL *et al.*, 2010] or `ABySS` [BIROL *et al.*, 2009] (not shown in **Figure 1.6**). Within the mapped dataset, TSSs are detected and based thereon, gene start sites can be corrected, ncRNAs identified, and novel genes, which were not predicted by the annotation software, annotated.

The second workflow (**Fig. 1.6B**) begins with a cDNA library that contains random fragments (not 5'-enriched) in order to measure expression levels of genes across their complete length. Sequencing and mapping of the library is performed in the same manner as in the first workflow. The reads that overlap a CDS are then counted and result in a table of read counts per gene for each experimental condition. Next, the raw read counts are normalized across all conditions in order to account for the varying efficiency of library preparations and sequencings. Thereafter, the differential expression testing takes place, which is preferentially carried out by software packages of the statistics language `R`, for instance `DESeq` [ANDERS & HUBER, 2010], `edgerR` [ROBINSON *et al.*, 2010] or `baySeq` [HARDCASTLE & KELLY, 2010]. The outcome of this procedure is a list of differentially expressed gene with associated fold-changes and p-values. These may be further analyzed for the identification of significantly differentially expressed genes and gene clusters. Moreover, gene sets derived from this analysis can be piped into *pathway enrichment tests* which then compute the likelihood of a given pathway to be significantly affected by the input genes [OSHLACK *et al.*, 2010] (not shown here).

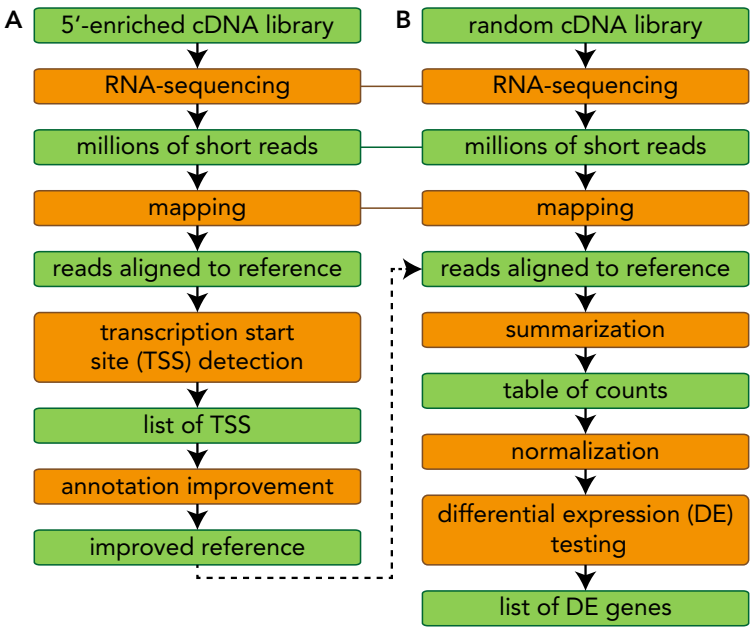

Figure 1.6.: RNA-seq workflows for (**A**) annotation improvement and (**B**) differential expression testing. The enhanced annotation can be used for as an improved reference during differential expression testing (dashed arrow).

1.9. Motivation and aims of this thesis

Acarbose is a potent drug for the treatment of diabetes mellitus type-2, an epidemic which claims more than 3 million victims every year [Roglic & Unwin, 2010]. Moreover, the incidence of diabetes is rapidly rising, which calls on the pharmaceutical industry to expand current production facilities or enhance their productivity in order to maintain access to affordable medical treatment [Whiting *et al.*, 2011]. In this regard, enhancing the productivity of the bacterial producer strains can be accomplished in several ways. Perhaps the most intuitive way is to optimize fermentation parameters and media compositions, followed by classical mutagenesis experiments with subsequent selection of higher-yielding strains. Although Bayer HealthCare AG followed this approach for many years very successfully, this strategy seems to have reached its limits by now and is generally superseded by modern genetic engineering approaches [Schedel, 2006]. As a prerequisite for targeted genetic modifications, the preferably complete genome sequence of the organism has to be known. For this reason it was decided to initiate the *Actinoplanes* genome project as a collaborative effort of the Center for Biotechnology, Bielefeld University and Bayer HealthCare AG, Wuppertal, in order to determine the genome sequence of an acarbose producer strain. The natural variant *Actinoplanes* sp. SE50/110 was selected because of its elevated, well measureable acarbose production of up to 1 g/L and because of its publicity in the scientific literature [Crueger *et al.*, 1998a, Stratmann *et al.*, 1999, Hemker *et al.*, 2001, Zhang *et al.*, 2002, Zhang *et al.*, 2003].

Assuming the successful genome sequencing and assembly, further work comprised the complete annotation of the genome as well as the analysis of special genes and gene clusters of interest. As scarcely anything was known about the genome sequence apart from the well studied acarbose gene cluster, its putative influence on acarbose production efficiency through e.g. nutrient uptake mechanisms or competitive secondary metabolite gene clusters was of foremost interest.

For measuring gene expression levels, the novel RNA-seq technology had to be established for this high-GC organism. First experiments should result in the identification of highly abundant transcripts and the putative functions of their encoding proteins within the cell. Particular attention should be given to the acarbose biosynthesis gene cluster and its putative regulation. Moreover, an assessment of the currently available software for RNA-seq analyses should be performed in order to reveal necessary future developments which are currently insufficiently addressed by the bioinformatics community.

Likewise, combined genomics and transcriptomics data should be used to address currently open scientific questions on the proposed acarbose biosynthesis and its metabolism in *Actinoplanes* sp. SE50/110.

In conclusion, the major aims of this study were:

1. to establish the complete genome sequence of *Actinoplanes* sp. SE50/110,
2. to infer an annotation for all genomic features,
3. to analyze special genes and gene clusters of interest, and
4. to evaluate transcriptomics experiments.

Chapter 2.
2 Materials and Methods

This chapter describes the details of the biotechnological and computational methods that were used in the course of this work. The sections are arranged in the logical order of the experiments that have taken place, beginning with the materials and methods used for genome DNA-sequencing of *Actinoplanes* sp. SE50/110. Following up, details about genome assembly and finishing procedures are provided. These are complemented by descriptions of bioinformatic tools employed for genome annotation and more advanced inquiries. At last, RNA-sequencing protocols and computational analysis strategies are outlined.

2.1. Acquisition of the strain *Actinoplanes* sp. SE50/110

A liquid culture of the strain *Actinoplanes* sp. SE50/110 was kindly provided by Bayer HealthCare AG, Wuppertal, Germany. The identical strain can also be obtained from the 'American Type Culture Collection' (P.O. Box 1549 Manassas, Virginia 20108, USA) via the identification number ATCC:31044 or the 'Centraalbureau voor Schimmelcultures' (P.O. Box 85167, 3508 AD Utrecht, Netherlands) via the identification number CBS:674.73.

2.2. Genomic DNA-sequencing methods

2.2.1. Cultivation of *Actinoplanes* sp. SE50/110 for DNA-sequencing

The *Actinoplanes* sp. SE50/110 strain was cultivated in a two-step shake flask system with the aim of subsequent DNA isolation. Beside inorganic salts the NBS medium (Bayer HealthCare AG, Wuppertal) contained starch hydrolysate as carbon source and yeast extract as nitrogen source (**Tab. 2.1**). Pre culture and main culture were incubated for 3 and 4 days, respectively, on a rotary shaker at 28 °C. Then the biomass was collected by centrifugation.

Table 2.1.: Components of the NBS medium used for cultivation of strains for genomic DNA isolation.

Component	Concentration
Starch hydrolysate	120.0 g/L
Asparagine × H_2O	15.0 g/L
Yeast extract	2.0 g/L
Tri-Natriumcitrate	7.5 g/L
$MgSO_4$ × 7 H_2O	2.0 g/L
$CaCl_2$ × 2 H_2O	2.0 g/L
$FeCl_3$ × 6 H_2O	0.5 g/L
K_2HPO_4	0.5 g/L
KH_2PO_4	0.5 g/L
in 1000 mL aqua dest.	

2.2.2. Isolation of genomic DNA from *Actinoplanes* sp. SE50/110

The preparation of genomic DNA of *Actinoplanes* sp. SE50/110 was performed as follows. The mycelium of 50 mL freshly grown culture was harvested by centrifugation (10 min, 3.350 rcf, 4 °C). The pellet was washed 4 times in a buffer containing 15% sucrose (Merck KGaA, Darmstadt, Germany), 25 mM Tris/HCl pH 7.2 (Merck KGaA, Darmstadt, Germany), and 25 mM EDTA (Merck KGaA, Darmstadt, Germany) under the same conditions. Finally the pellet was resuspended in 4.5 mL of the same buffer and lysozyme (Merck KGaA, Darmstadt, Germany) and RNAse (Qiagen, Hilden, Germany) were added to final concentrations of 5 mg/mL and 50 µg/mL respectively and the mixture was incubated at 37 °C for 45 min. After the addition of SDS (Serva, Heidelberg, Germany) and proteinase K (Qiagen, Hilden, Germany) to 0.1% and 2 µg/mL final concentrations, respectively, the incubation was continued at 50 °C for 5 min. NaCl (Merck KGaA, Darmstadt, Germany) was added to a final concentration of 300 mM and the volume adjusted with water for injection (highly pure water) to 8 mL. The lysate was subjected to three successive phenol/SEVAG extractions (SEVAG is a mixture of 24 parts chloroform [Merck KGaA, Darmstadt, Germany] and 1 part isoamylalcohol [Merck KGaA, Darmstadt, Germany]) and the phenol was removed by washing the DNA solution with 10 mL SEVAG. The DNA was precipitated by the addition of 0.1 volume of 3 M sodium acetate (pH 4.8) (Merck KGaA, Darmstadt, Germany) and 1 volume of cold isopropanol (Merck KGaA, Darmstadt, Germany). The DNA was pelleted by centrifugation (25 min, 3.350 rcf, 4 °C), the DNA pellet was washed 5 times with 70% ethanol (Merck KGaA, Darmstadt, Germany) (10 min, 3.350 rcf, 4 °C) and air-dried. Finally the pellet was resuspended in 400 µL Tris buffer pH 8.5 over night at 4 °C and the DNA concentration was determined by measuring the optical density at 260 nm and 280 nm. The correct size of the prepared DNA was analyzed by agarose gel electrophoresis.

2.2.3. Pyrosequencing of the *Actinoplanes* sp. SE50/110 genomic DNA on the Genome Sequencer FLX

The Genome Sequencer (GS) FLX system (454 Life Sciences) has been used for pyrosequencing of *Actinoplanes* sp. SE50/110. Two different library preparation and sequencing protocols as well as sequencing chemistries were used on the GS FLX platform:

- Standard sequencing chemistry with long PE protocol. The DNA fragment size for the PE library construction was $2,787$ bases $\pm$ 696. The protocol yielded an average read length of 251 bases and a total number of 259 Mb was sequenced in two full runs.

- Titanium sequencing chemistry with WGS protocol. The DNA fragment size for the WGS library construction was 500-800 bp. The protocol yielded an average read length of 537 bases and a total number of 198 Mb was obtained from a half picotiter plate. For the emulsion PCR (emPCR), 1.5 mL H_2O was substituted for an equal amount of emPCR Additive kindly provided by 454 Life Sciences.

Details on the protocols are provided in the corresponding manufacturer's manuals, namely the 'GS FLX Sequencing Method Manual' (December 2007), 'GS FLX Paired End DNA Library Preparation Method Manual' (December 2007), 'GS FLX Titanium Sequencing Method Manual' (October 2008) and the 'GS FLX Titanium General Library Preparation Method Manual' (October 2008).

2.3. Genome assembly and mapping techniques

2.3.1. Genome assembly

The automated assembly of all *Actinoplanes* sp. SE50/110 reads generated by the GS FLX platform was performed with the `Newbler` assembler software (`gsAssembler` version 2.0.00.22, 454 Life Sciences). For the assembly process, standard settings were used. For detailed information on the assembly algorithm see the Genome Sequencer FLX System Software Manual, version 2.0.

2.3.2. Read mapping on the acarbose gene cluster

The mapping of single reads and contigs on the acarbose gene cluster reference sequence [GenBank:Y18523.4] has been carried out with the `BLAST` software package [Altschul *et al.*, 1990]. Subsequent parsing of the results, calculation of GC-content and visualizations were implemented by custom perl scripts developed for this purpose at the Center for Biotechnology, Bielefeld University, Germany.

2.4. Genome finishing methods

2.4.1. Construction of a fosmid library for the *Actinoplanes* sp. SE50/110 genome finishing

The fosmid library construction for *Actinoplanes* sp. SE50/110 with an average insert size of 40 kb has been carried out on isolated genomic DNA by IIT Biotech GmbH (Universitätsstrasse 25, 33615 Bielefeld, Germany). For construction in *Escherichia coli* EPI300 cells, the CopyControl™ Cloning System (EPICENTRE Biotechnologies, 726 Post Road, Madison, WI 53713, USA) has been used. The kit was obtained from Biozym Scientific GmbH (Steinbrinksweg 27, 31840 Hessisch Oldendorf, Germany).

2.4.2. Polymerase chain reactions

The gap sequences between the contigs of the *Actinoplanes* sp. SE50/110 genome were amplified by PCRs using a Mastercycler pro S (Eppendorf) device. PCRs were performed using the Phusion High-Fidelity Master Mix (Finnzymes) in conjunction with varying combinations and concentrations of the anti-self annealing additives DMSO, betaine, and trehalose. Device run protocols were adapted in temperature, cycle number and speed according to the expected gap length and primer composition. Primers were ordered from Metabion AG after selection by using program features of the `Consed` software [Gordon *et al.*, 1998, Gordon *et al.*, 2001] or a custom implemented perl script. The PCR products were controlled for single bands in 2% agarose gels prior to sequencing.

2.4.3. Sanger sequencing of PCR products and terminal insert sequences from the *Actinoplanes* sp. SE50/110 fosmid library

The fosmid library terminal insert sequencing was carried out with capillary sequencing technique on a 3730xl DNA-Analyzer (Applied Biosystems) by IIT Biotech GmbH, Bielefeld, Germany. The resulting chromatogram files were base called using the `phred` software [Ewing *et al.*, 1998, Ewing & Green, 1998] and stored in FASTA format. Both files were later used for gap closure and quality assessment using the `Consed` software [Gordon *et al.*, 1998, Gordon *et al.*, 2001].

2.4.4. Finishing of the *Actinoplanes* sp. SE50/110 genome sequence by manual assembly

In order to close remaining gaps between the contigs that were still present after the automated assembly, the visual assembly software package `Consed` [Gordon *et al.*, 1998, Gordon *et al.*, 2001] was utilized. Within the graphical user interface, fosmid walking primer and genome PCR primer pairs were selected at the ends of contiguous contigs. These were used to amplify desired sequences from fosmids or genomic DNA in order to bridge the gaps between contiguous contigs.

After the DNA sequence of these amplicons had been determined, manual assembly of all applicable reads was performed with the aid of different `Consed` program features. In cases where the length or quality of one read was not sufficient to span a gap, multiple rounds of primer selection, amplicon generation, amplicon sequencing, and manual assembly were performed.

2.5. Computational genome annotation

2.5.1. Prediction of coding sequences on the *Actinoplanes* sp. SE50/110 genome sequence

The potential genes were identified by a series of programs which are all part of the `GenDB` annotation pipeline [Meyer *et al.*, 2003]. For the automated identification of CDSs, the prokaryotic gene finders `Prodigal` [Hyatt *et al.*, 2010] and `GISMO` [Krause *et al.*, 2007] were primarily used. In order to optimize results and allow for easy manual curation, further intrinsic, extrinsic, and combined methods were applied by means of the `Reganor` software [McHardy *et al.*, 2004, Linke *et al.*, 2006] which utilizes the popular gene prediction tools `Glimmer` [Salzberg *et al.*, 1998, Delcher *et al.*, 1999] and `CRITICA` [Badger & Olsen, 1999].

2.5.2. Functional annotation of the identified CDS on the *Actinoplanes* sp. SE50/110 genome

The identified CDSs were analyzed through a variety of different software packages in order to draw conclusions from their DNA- and/or amino acid-sequences regarding their potential function. Besides functional predictions, further characteristics and structural features have also been calculated.

Similarity-based searches were applied to identify conserved sequences by means of comparison to public and/or proprietary nucleotide- and protein-databases. If a significant sequence similarity was found throughout the major section of a gene, it was concluded that the gene should have a similar function in *Actinoplanes* sp. SE50/110. The similarity-based methods, which were used to annotate the list of ORFs, are termed `BLASTP` [Coulson, 1994] and `RPS-BLAST` [Marchler-Bauer *et al.*, 2002].

Enzymatic classification has been performed on the basis of enzyme commission (EC) numbers [NC-ICBMB & Webb, 1992, Bairoch, 2000]. These were primarily derived from the PRIAM database [Claudel-Renard *et al.*, 2003] using the `PRIAM_Search` utility on the latest version (May 2011) of the database. For genes with no PRIAM hit, secondary EC-number annotations were derived from searches against the Kyoto encyclopedia of genes and genomes (KEGG) database [Kanehisa *et al.*, 2006, Kanehisa & Goto, 2000]. For further functional gene annotation, the cluster of orthologous groups of proteins (COG) classification system has been applied [Tatusov *et al.*, 1997, Tatusov *et al.*, 2001] using the latest version (March 2003) of the database [Tatusov *et al.*, 2003].

To identify potential transmembrane proteins, the software `TMHMM` [Sonnhammer *et al.*, 1998, Krogh *et al.*, 2001] has been utilized.

The software `SignalP` [Nielsen *et al.*, 1997, Nielsen & Krogh, 1998, Bendtsen *et al.*, 2004] was used to predict the secretion capability of the identified proteins. This is done by means of hidden Markov models (HMMs) and neural networks, searching for the appearance and position of potential signal peptide cleavage sites within the amino acid sequence. The resulting score can be interpreted as a probability measure for the secretion of the translated protein. `SignalP` retrieves only those proteins which are secreted by the classical signal-peptide-bound mechanisms.

In order to identify further *Actinoplanes* sp. SE50/110 proteins which are not secreted via the classical way, the software `SecretomeP` has been applied [Bendtsen *et al.*, 2005]. The underlying neural network has been trained with secreted proteins, known to lack signal peptides despite their occurrence in the exoproteome. The final secretion capability of the translated genes was derived by the combined results of `SignalP` and `SecretomeP` predictions.

To reveal polycistronic transcriptional units, proprietary software has been developed which predicts jointly transcribed genes by their orientation and proximity to neighboring genes (adopted from [Salgado *et al.*, 2000]). In light of these predictions, operon structures can be estimated and based upon them further sequence regions can be derived with high probability of contained promoter and operator elements.

The software `DNA mfold` [Zuker, 2003] has been used to calculate hybridization energies for the DNA sequences in order to identify secondary structures such as transcriptional terminators which indicate operon and gene ends, respectively. The involved algorithm computes the most stable secondary structure of the input sequence by striving to the lowest level in terms of Gibbs free energy (ΔG), a measure for the energy which is released by the formation of hydrogen bonds between the hybridizing base pairs.

2.5.3. Phylogenetic analyses

The 16S rDNA-based phylogenetic analysis was performed on DNA sequences retrieved by public `BLAST` [Altschul *et al.*, 1990] searches against the 16S rDNA of *Actinoplanes* sp. SE50/110. From the best hits, a multiple sequence alignment was built by applying the `MUSCLE` program [Edgar, 2004b, Edgar, 2004a] before deriving the phylogeny thereof by the `MEGA 5` software [Tamura *et al.*, 2007] using the neighbor-joining algorithm [Saitou & Nei, 1987] with Jukes-Cantor model [Jukes & Cantor, 1969]. Genome based phylogenetic analysis of *Actinoplanes* sp. SE50/110 in relation to other closely related species was performed with the comparative genomics tool `EDGAR` [Blom *et al.*, 2009]. Briefly, the core genome of all selected strains was calculated and based thereon, phylogenetic distances were calculated from multiple sequence alignments. Then, phylogenetic trees were generated from concatenated core gene alignments.

2.6. RNA-sequencing and analysis

2.6.1. Cultivation of *Actinoplanes* sp. SE50/110 for RNA-sequencing

Actinoplanes sp. SE50/110 pre-cultures were inoculated with 3.5 mL glycerin cryo cultures and cultivated in 500 mL baffled polycarbonate Erlenmeyer flasks (Corning) with Silicosen C-55 plugs (Hirschmann Laborgeräte) in 100 mL modified NBS medium (Bayer HealthCare AG, Wuppertal) for 5 days at 140 rpm and 28 °C in a GFL shaking incubator 3032 (GFL). The modified NBS medium is a glucose containing complex medium (Glc-CM), whose exact composition is listed in **Table 2.2**. For the inoculation of main cultures, pre-cultures were centrifuged at 3.500 rcf for 3 min and washed twice with 50 mL of sterile 150 mM NaCl solution. After another centrifugation step (2.250 rcf, 3 min), the supernatant was decanted. The resulting pellet was resuspended in 30 mL NaCl solution of which 2 mL were used for inoculation of main cultures.

The main cultures were grown in 250 mL baffled polycarbonate Erlenmeyer flasks (Corning) with Silicosen C-40 plugs (Hirschmann Laborgeräte) in 50 mL medium for 4 days at 140 rpm and 28 °C in a GFL shaking incubator 3032 (GFL). Three different media were used for the cultivation of *Actinoplanes* sp. SE50/110. First, a maltose containing minimal medium (Mal-MM) in four replicates (**Tab. 2.3**); second, the same medium with an additionally supplemented trace element solution (Mal-MM-TE) in four replicates (**Tab. 2.4**); and third, the glucose containing complex medium (Glc-CM) in two replicates (**Tab. 2.2**), which was also used for the pre-cultivation. The Mal-MM medium is based on the Cerestar medium [Rockser & Wehmeier, 2009], but was changed regarding several aspects. It consists of four instead of three solutions, lacks yeast extract as a complex component, and includes new compounds and adjusted concentrations. During its preparation, solutions 1-3 were combined and sterile filtered, while solution 4 was sterile filtered separately and added afterwards. For Mal-MM-TE, 1 mL of sterile filtered 1:100 trace element solution was added to Mal-MM.

Table 2.2.: Components of the glucose complex medium (Glc-CM).

Component	Concentration
D-glucose × 1 H_2O	11.00 g/L
casein peptone	4.00 g/L
yeast extract	4.00 g/L
$MgSO_4$ × 7 H_2O	0.50 g/L
K_2HPO_4 × 3 H2O	2.00 g/L
KH_2PO_4	4.00 g/L
pH adjusted to 7.3 with NaOH	

Table 2.3.: Components of the maltose minimal medium (Mal-MM).

Solution	Component	Amount
solution 1	maltose × 1 H_2O (204 mM)	73.68 g
	$(NH_4)_2SO_4$	5.00 g
	casamino acids	3.00 g
	in 400 mL aqua dest.	
solution 2	K_2HPO_4 × 3 H_2O	6.55 g
	KH_2PO_4	5.00 g
	in 300 mL aqua dest.	
solution 3	FeCl2 × 4 H_2O (460 mg/mL)	184 mg
	trisodium citrate × 2 H_2O	5.70 g
	in 300 mL aqua dest.	
solution 4	$MgCl_2$ × 6 H_2O	1.00 g
	$CaCl_2$ × 2 H_2O	2.00 g
	in 20 mL aqua dest.	

Table 2.4.: Trace elements solution for maltose minimal medium with trace elements (Mal-MM-TE).

Trace element	Component	Concentration
zinc	$ZnCl_2$	40 mg/L
iron	$FeCl_3$ × 6 H_2O	200 mg/L
copper	$CuCl_2$ × 2 H_2O	10 mg/L
manganese	$MnCl_2$ × 4 H_2O	10 mg/L
sodium & boron	$Na_2B_40_7$ × 10 H_2O	10 mg/L
molybdenum	$(NH_4)_6Mo_7O_24$ × 4 H_2O	10 mg/L
supplemented 1:100 with Mal-MM to yield Mal-MM-TE		

2.6.2. Total RNA isolation from *Actinoplanes* sp. SE50/110

For the total RNA isolation of all biological replicates (4× Mal-MM, 4× Mal-MM-TE, and 2× Glc-CM), several 1.5 mL aliquots of each cultivation flask were centrifuged (3 sec at 20,000 rcf) and the supernatant discarded. The resulting pellets were frozen in liquid nitrogen and stored at -80 °C. The total RNA was then prepared using two different methods: TRIzol reagent (Life Technologies) and a commercial kit using spin columns (RNeasy Mini Kit, QIAGEN). For each method two pellets from one biological replicate were pooled in the initial resuspension step.

For the first-mentioned method, frozen cell pellets were resuspended in 1 mL TRIzol, immediately transferred into tubes containing silica beads (Lying Matrix B, MP Biomedicals) and homogenated (RiboLyser, Hybaid) for two cycles (20 sec at

speed 6.5) with cooling on ice for 1 min in between. The samples were then centrifuged at 13,000 rcf and 4 °C for 3 min. The resulting supernatant was transferred to tubes containing 200 µL of chloroform. The mixture was vortexed, incubated for 1 min at room temperature and centrifuged at 10,000 rcf for 10 min. Following phase separation, the aqueous phase was removed and pipetted into a separate reaction tube. After adding 450 µL isopropanol, the mixture was incubated on ice for 10 min and then centrifuged at 16,000 rcf and 4 °C for 15 min. The resulting pellet was washed with 75% ethanol, dried at 37 °C, and resuspended in 122.5 µL DEPC-H_2O. The samples were treated with DNase I (Roche) and incubated for 30 min at 37 °C. Afterwards, 150 µL phenol/chloroform/isoamyl alcohol (25:24:1, ROTH) was added, the mixture vortexed, and centrifuged at 20,000 rcf for 15 min. The aqueous phase was removed and pipetted into a separate tube. After adding 450 µL 0.3 M sodium acetate, the mixture was centrifuged at 16,000 rcf and 4 °C for 20 min. The precipitated nucleic acids were washed twice with 75% ethanol, dried at 37 °C, resuspended in 124 µL DEPC-H_2O and another DNase I treatment (QIAGEN) was carried out. Finally, the isolated RNA was stored at -80 °C.

For the second-mentioned method, RNA isolation was performed using the RNeasy Mini Kit (QIAGEN). Frozen cell pellets were resuspended in 800 µL of RLT buffer immediately transferred into Lying Matrix tubes and homogenated for two cycles (20 sec at speed 6.5) with cooling on ice for 1 min in between. On-column digestion of chromosomal DNA was performed with DNase I (QIAGEN). To completely remove all remaining DNA traces, a second digestion was performed with DNase I from Roche. Subsequently, the RNA was purified according to the manufacturer's instructions.

The quality of the isolated total RNA was assessed by PCR with gene-specific primers and with capillary gel electrophoresis (Agilent Bioanalyzer 2100, Agilent Technologies), which reported the RNA integrity numbers 5.5 for Mal-MM, 5.2 for Mal-MM-TE, and 3.6 for Glc-CM. The corresponding electropherograms are shown in **Figure A.2**.

2.6.3. Preparation of cDNA libraries and high-throughput sequencing

As a first step in preparation of the sequencing libraries, duplicate RNA samples (5 µg portions each) were processed with the RiboZero rRNA removal kit (Gram-Positive; Epicentre) as recommended by the manufacturer to deplete the amounts of 23S, 16S, and 5S rRNAs, respectively. For further processing, the two samples were pooled. Next, the RNA was fragmented to a size of 200-500 bp by metal hydrolysis, using 6 mM MgOAc and 20 mM KOAc for 150 sec at 94 °C, and the reaction was stopped by addition of an equal volume of ice-cold buffer and incubated at 0 °C for 5 min. After precipitation using NaAc and ethanol, the RNA was resuspended in a volume of 37 µL DEPC-H_2O and 2 µL were used for size control in the Agilent Bioanalyzer using the RNA Pico Kit. The 5'-tri- and diphosphorylated ends were converted to 5'-monophosphates by RNA polyphosphatase (Epicentre) and unphosphorylated ends were 5'-monophosphorylated using polynucleotide kinase (NEB) as recommended by the supplier. The sequencing libraries were generated with the help of the TruSeq RNA

sample prep Kit v.2 (Illumina) and sequencing of the 26 bases runs were performed using the Genome Analyzer IIx machine (Illumina).

The cDNA library preparation for 5'-enriched transcripts were performed as above with an additional digestion of stabile RNAs using terminator exonuclease (TEN) after application of the RiboZero rRNA removal kit. Sequencing was then performed as single read, 36 bases on the same device.

2.6.4. Determination of cell dry weights of *Actinoplanes* cultures

The cultures were harvested by centrifugation (2 min, 4000 rfc) in 50 mL centrifuge tubes (Greiner Bio-One) and washed twice with washing buffer (50 mM Tris/HCl pH 7.2). The resulting pellets were freeze-dried and weighed.

2.6.5. Quantification of acarbose in the supernatant of *Actinoplanes* cultures by HPLC and UV detection

For acarbose quantification using high performance liquid chromatography (HPLC), 10 µL of sterile filtered supernatant were quantified by HPLC (KNAUER, Smartline Manager 5000, Smartline Pump 1000, UV Smartline Detector 2500 and Spark Holland BV, Triathlon autosampler). An isocratic flow of 1.7 mL/ min of 64% acetonitrile, 10% methanol and 26% phosphate buffer consisting of 0.62 g/L KH_2PO_4 and 0.38 g/L $K_2HPO_4 \times 2\ H_2O$ was applied over an Hypersil APS-2 precolumn cartridge (MZ Analysentechnik, No. VK 5.4, 0.6085) and a Hypersil APS-2 analytical column (Thermo Scientific, No. 30703-124030). The temperature was adjusted to 33 °C. Acarbose was detected at 210 nm against an acarbose standard (Mat. Nr: 05479894, Charge: BXR2UBZ) kindly provided by Bayer HealthCare AG.

2.6.6. Bioinformatic analysis of RNA-seq results

Read mapping The sequenced reads were quality filtered and subsequently mapped to the *Actinoplanes* sp. SE50/110 genome [GenBank:CP003170] using the exact mapping algorithm `SARUMAN` [Blom *et al.*, 2011]. The software reports all genome wide matching positions of a given read, which allows for rapid identification and filtering of reads matching to repetitive regions of the genome. Ambiguously mapping reads were excluded from the analysis. This stringent approach holds further benefits by excluding reads stemming from the six ribosomal DNA operons of *Actinoplanes* sp. SE50/110. To account for possible sequencing errors, a single mismatch was allowed to occur in the alignment of a read in case no perfect match has been found previously.

Detection of transcription start sites For the reason that no suitable public software was available, the detection of TSSs was performed by custom programs implemented in the perl programming language. The developed algorithm performed three major tasks:

1. Read-mapping results from `SARUMAN` are converted into forward and reverse strand histograms of the reference genome, resulting in a base-specific resolution of coverage (reads per base) for the *Actinoplanes* sp. SE50/110 genome.
2. TSSs are detected on the two histograms by comparing the differences in coverage between two adjacent bases while traversing the complete histograms. If the difference (Δ stacksize) exceeds a user specified threshold, the base position is reported as a putative TSS. In order to account for possible microheterogeneity, i.e. multiple TSSs in close proximity [Knippers, 2001], no new TSS are detected within *read length* after the previously reported TSS on the same strand.
3. Additional information are calculated for every candidate TSS regarding its next upstream, downstream, and overlapping gene (if any) on the sense and antisense strands. In particular, the gene identifier, gene type (rRNA, tRNA, ncRNA, protein coding gene), TSS distance to CDS start, and TSS distance to CDS stop are reported along with a user defined upstream and/or downstream DNA region surrounding the TSSs.

Given these information, research question specific filtering of the data was performed in spreadsheet applications.

Read summarization For each gene, all reads that overlapped its coding region in the sense direction were counted and the sum assigned to the gene as its raw read count. Additionally, also antisense read counts were determined for each gene. These steps were implemented by a custom perl script because public software, such as the `HTSeq-count` module of EMBL's `HTSeq`[1], does not support antisense determination.

Normalization In order to account for differences in the number of sequenced and mapped reads between the three media conditions, which were caused by varying RNA isolation- and cDNA library preparation efficiencies, a normalization procedure was applied. In detail, from the total number of reads that overlapped coding sequences in each condition global scaling factors were calculated. For subsequent differential expression testing, these factors were passed to the corresponding program (see next paragraph). For visualization and detection of most abundant transcripts however, the factors were used in the calculation of reads per kilobase of coding sequence per million mapped reads (RPKM), as given in the following formula.

$$RPKM_{c,g} = \frac{overlapping\ reads_{c,g}}{length_g} \times \frac{1000000}{mapped\ reads_c}$$

where c is the condition and g the gene of interest. The first term normalizes the length of each gene, whereas the second term accounts for the differences in library size, yielding a gene-length- and library-size-independent expression measure [Mortazavi *et al.*, 2008].

[1] `HTSeq` is available at http://www-huber.embl.de/users/anders/HTSeq

Differential expression testing The software `DESeq` [Anders & Huber, 2010] was used in this work because it allowed for differential expression testing of experiments with no replicates. Furthermore, it estimates the parameters for the underlying negative binomial distribution from all biological conditions available, thereby assuming that there is no true differential expression for most of the genes. This is in agreement with the expectations and implies rather conservative results which seem to be appropriate for experimental setups with no replicates. Besides library-size normalized read counts per gene, `DESeq` reports the fold-change and p-value of a differential expression test for pairwise comparison of conditions.

2.7. Gas-chromatographic analysis of the anti-self-annealing additive

The derivatization and gas-chromatography mass-spectrometry (GC-MS) analysis was performed as previously published [Watt *et al.*, 2009]. The substance with inhibitory effects to self-annealing sequences was kindly provided by 454 Life Sciences and analyzed after 1:600 dilution.

3 Chapter 3. Results

This chapter comprises the outcomes of the conducted experiments that were performed during this work. The sections are arranged in temporal relation to the analyses that were carried out, starting with the identification and remedy of obstacles encountered during the initial genome sequencing of *Actinoplanes* sp. SE50/110. Thereafter, results from the assembly and finishing phases are reported that lead to the reconstruction of the complete genome sequence – the most labour-intensive part of this project. Next, genome annotation outcomes are described and more advanced analyses demonstrated. The last part reports the findings from conducted transcriptomics experiments, which on the one hand, helped to improve the genome annotation and, on the other hand, allowed for differential expression testing between different cultivation conditions of *Actinoplanes* sp. SE50/110.

3.1. Solving the high-GC problem for *Actinoplanes* sp. SE50/110 genome sequencing

The initial genome sequencing of *Actinoplanes* sp. SE50/110 was carried out with two full runs on the Genome Sequencer FLX using standard chemistry and the paired-end protocol. In each run about 100 million bases were sequenced, yielding approximately 750,000 shotgun reads and 260,000 PE reads (**Tab. 3.1**).

Although the joint assembly of both PE runs by the `Newbler` software was successful with more than 99.5% of assembled reads and a low inferred read error of less than 0.4% (**Tab. 3.2**), it resulted in a disillusioning genome wide assembly of 7,973 (4,358 $\geq$ 500 b) contigs. The preliminary genome size was found to be 8.33 MB (**Tab. 3.3**). Based on the number of aligned bases and the genome size, a mean coverage of 24.58 reads per base was determined, whereas the average genome GC-content was calculated to be 70.44%. Due to the use of a PE library, scaffolding information constituted 307 scaffolds, containing 7.87 MB of the genome sequence.

Table 3.1.: Results of all three sequencing runs.

Run	454 Technology	Reads	Paired reads	Bases
1	Standard, PE	742,169	259,260	103,840,588
2	Standard, PE	751,570	265,457	105,329,378
3	Titanium, WGS	481,602	-	197,732,895
	Total	1,975,341	524,717	406,902,861

Table 3.2.: Results of successfully assembled reads, bases and the inferred read errors.

Run	454 Technology	Assembled reads	Assembled bases	Inferred read error[a]
1	Standard, PE	739,079 (99.58%)	101,847,643 (98.08%)	370,520 (0.36%)
2	Standard, PE	748,526 (99.59%)	103,411,267 (98.18%)	364,397 (0.35%)
3	Titanium, WGS	480,863 (99.85%)	196,416,109 (99.33%)	1,018,256 (0.52%)
	Total	1,968,468 (99.65%)	401,675,019 (98.72%)	1,753,173 (0.44%)

[a]The inferred read error is calculated from mismatches between the reads and the consensus sequence of the final assembled contigs and measures the frequency of incorrectly called bases.

3.1.1. Analysis of gap regions resulted from standard PE sequencing

In order to reveal the reasons for the unusually high number of contigs generated by the `Newbler` assembly software, several potential causes of failure were analyzed. Starting with the analysis of the read data, `BLAST` [Altschul *et al.*, 1990] searches against public databases were performed in order to identify possible contaminations of the starting DNA. However, these did not provide indications for undesired DNA in the setup. Subsequently, it was investigated if reads were not correctly assembled by the assembly software and, as a consequence, gaps may have been introduced. For this analysis, a previously sequenced and assembled reference sequence has been utilized. The 41,323 bases long sequence hosts the acarbose biosynthetic gene cluster [GenBank:Y18523.4] which has previously been determined and analyzed [Wehmeier & Piepersberg, 2004]. Using this gene cluster as a reference sequence, the mapping of the contigs from the genome sequencing assembly resulted in the alignment of 30 contigs (**Fig. 3.1**). To rule out false alignments and possible problems resulting from repetitive elements, the 30 contigs were checked by bidirectional `BLAST` comparisons and found to be unique sequences, only occurring once within the acarbose cluster. In addition, the assembly process as a possible source of error could also be excluded because a mapping of all reads against the reference sequence showed uncovered regions exactly at the boundaries of the assembled contigs (**Fig. 3.2D**). These findings provided indication for the problem being involved in earlier steps of the sequencing process, as the regions between adjacent contigs could obviously not be sequenced although the average coverage was reasonably high. From the large amount of un-

Table 3.3.: Assembly results of standard PE and Titanium WGS sequencing runs for *Actinoplanes* sp. SE50/110.

Sequencing property	Standard chemistry, PE library, without emPCR Additive	Titanium chemistry, WGS library, with emPCR Additive
No. of reads	1,487,605	480,863
Percent of aligned reads	98.77	99.55
Percent of aligned bases	97.95	97.70
No. of all contigs	7,973	571
No. of bases in all contigs	8,333,840	9,091,694
No. of large contigs ($\geq$ 500 b)	4,358	510
No. of bases in large contigs	7,303,291	9,075,703
Percent of genome GC-content	70.44	71.31
Average genome coverage	24.58	21.25

covered regions it was furthermore concluded that the variance of the coverage was comparably high. As a matter of fact, the standard deviation has been calculated to be 16.69 for a mean coverage of 23.99 reads per base in the acarbose cluster.

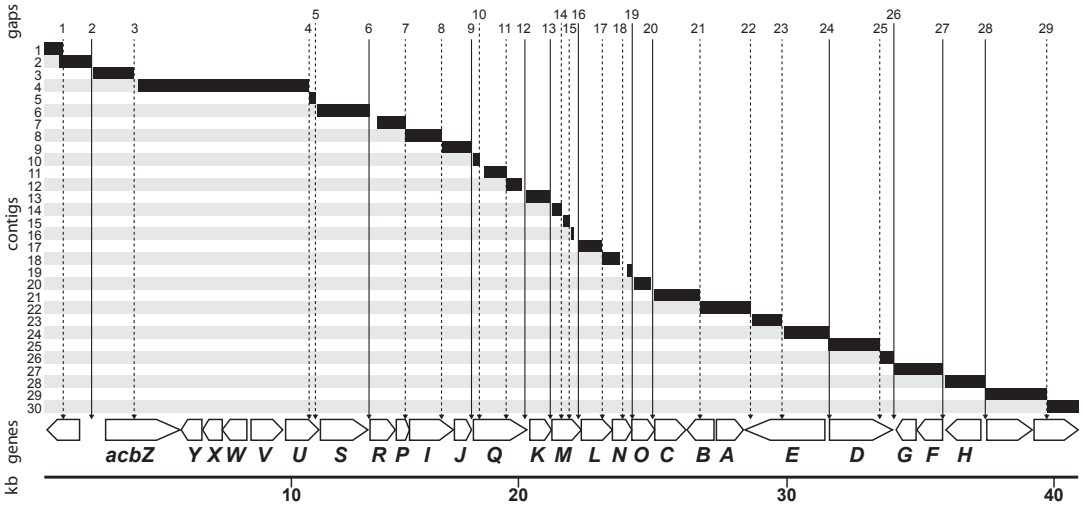

Figure 3.1.: The gaps and contigs of the *Actinoplanes* sp. SE50/110 acarbose gene cluster resulting from a standard paired end pyrosequencing run. The depicted 41 kb long acarbose biosynthesis gene cluster reference sequence hosts 28 genes which were previously named and annotated [Wehmeier & Piepersberg, 2004]. Constructed by the `Newbler` assembly software, the 30 contigs were subsequently mapped on the reference sequence using the `BLAST` [Altschul *et al.*, 1990] program. Between adjacent contigs overall 29 gaps occurred which can be further subdivided into 12 gaps that are located between adjacent genes (marked by solid lines) and 17 gaps which are located within coding sequences (marked by dashed lines).

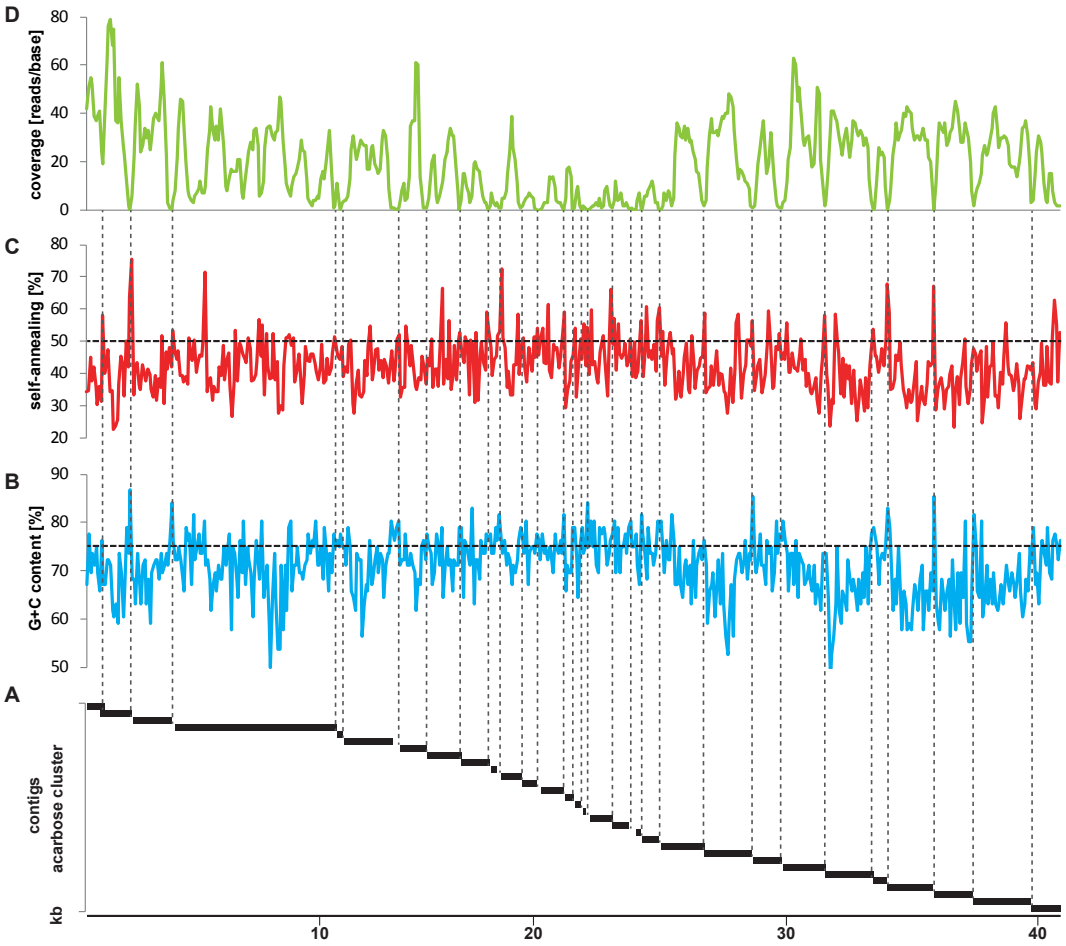

Figure 3.2.: The GC-content, the self-annealing energy and the observed coverage throughout the complete 41 kb long acarbose gene cluster of *Actinoplanes* sp. SE50/110. (**A**) The 30 contigs and 29 gaps resulting from the assembly of reads obtained from the standard paired end pyrosequencing run. The vertical dashed lines mark the gaps between adjacent contigs. (**B**) Changes in the GC-content of the acarbose gene cluster sequence. The horizontal dashed line marks 75% GC-content. (**C**) Self-annealing property of local 76 bases long sequence chunks throughout the complete gene cluster. The horizontal dashed line marks 50% hybridization, where 100% is defined as a perfect hairpin with 100% GC-content and 76 bases length. (**D**) Coverage in terms of reads per base aligned to the acarbose gene cluster sequence using the BLAST software [Altschul *et al.*, 1990].

The acarbose cluster is the longest and best studied contiguous sequence of the *Actinoplanes* sp. SE50/110 genome known to date [Wehmeier & Piepersberg, 2004]. It is therefore best suited to serve as the reference when mapping of contigs and reads had to be done. This is especially important as longer sequences generally reduce the possibility to deal with atypical sequence information like lytic bacteriophages, IS-elements or bacterial telomeres. Despite its short length in comparison to the genome

size, it is justifiable to convey structural results from the analysis of this cluster to the genome to a certain extent, as parameters such as GC-content (genome average: 70.44%; acarbose cluster: 70.81%), coverage (genome average: 24.58 fold; acarbose cluster: 23.99 fold) and contig length (genome average: 1.38 kb; acarbose cluster: 1.04 kb) are close to the genome averages. Furthermore, the gene organization forms a typical cluster for the production of secondary metabolites with several putative operons, which are also found in Streptomyces [Rockser & Wehmeier, 2009, Kaysser *et al.*, 2009] and other *Actinoplanes* species [Boakes *et al.*, 2009, Boakes *et al.*, 2010]. Therefore it can be assumed that findings will also be valid for most of the remaining genome.

3.1.2. The gaps in the *Actinoplanes* sp. SE50/110 acarbose gene cluster are due to an extremely low read coverage

In order to further investigate the inability of the sequencing to generate reasonable numbers of reads at the boundaries of the contigs or in gap regions respectively, bioinformatic analysis of the gaps and contigs located on the acarbose reference sequence were carried out. The mean gap size between adjacent contigs on the acarbose cluster was found to be 76 bases in length with an average GC-content of 78.6% (**Tab. 3.4**). In three of the 29 gaps, namely gap #21, #24, and #29, the adjacent contigs do actually overlap by 5, 20 and, 25 bases, respectively (indicated by negative numbers in the column *gap length*). However, as the overlaps were not supported by sufficient length and coverage, the assembly process did not join these contigs. Another interesting finding of the mapping procedure revealed that 12 of the 29 gaps are directly located in between adjacent genes (**Fig. 3.1**, solid lines), whereas the remaining 17 gaps are located within genes (**Fig. 3.1**, dashed lines). This suggests the involvement of rho-independent transcriptional terminators, as analyzed in **Section 3.1.3**.

Using `BLAST`, a more detailed analysis of the coverage has been performed by mapping of all reads onto the reference sequence. After applying a sliding window approach on these results (window length was 76 bases as determined above), a high correlation coefficient of -0.58 was found by comparing the read-coverage (average number of reads covering a certain nucleotide base in the 76 bases window) with the GC-content of these sequence chunks. In other words, in most cases where a high GC-content is observed, the coverage drops accordingly and vice versa. To investigate the impact of this correlation on the actual assembly, a direct comparison of GC-content and the read-coverage of the reference sequence was performed as it is known that a high GC-content may lead to sequencing problems [Frey *et al.*, 2008] and indeed, the negative correlation coefficient of -0.57 of the two parameters can also be clearly observed (**Fig. 3.2B&D**). In addition, it is shown that in practically every case where two contigs were interrupted by a gap, the coverage in this area drops close to zero reads per base (**Fig. 3.3**), proving the initial assumption that these sequences have not been sequenced and the gaps are not due to assembly errors. It was also found that the GC-content in gap regions always rises above 75% (**Fig. 3.2B**, horizontal dashed line). However, from these results it is also evident that not every high-GC stretch leads

Table 3.4.: Properties of the gaps and contigs from the *Actinoplanes* sp. SE50/110 acarbose gene cluster resulting from standard pyrosequencing.

Contigs				Gaps / overlaps					
#	Length [bp]	GC-content [%]		#	Gap length[a] [bp]	Gap GC-content [%]	Hairpin size [bp]	Hairpin GC-content [%]	Hairpin fold-back[b] [%]
1	719	69.16	⇐	1	1	100.00	48	83.33	58.20
2	1176	68.37	⇐	2	48	86.80	40	92.50	75.40
3	1641	69.53	⇐	3	135	84.20	40	85.00	53.60
4	6822	70.86	⇐	4	25	77.60	70	80.00	51.50
5	266	75.94	⇐	5	20	70.00	48	81.25	48.50
6	2110	70.33	⇐	6	325	80.30	42	85.71	52.00
7	1108	71.84	⇐	7	64	73.70	62	80.65	36.90
8	1410	71.41	⇐	8	13	78.90	56	87.50	52.90
9	1151	73.41	⇐	9	65	77.60	40	82.50	53.00
10	258	75.58	⇐	10	234	76.30	46	78.26	72.40
11	842	72.68	⇐	11	16	76.30	70	78.57	39.70
12	605	74.21	⇐	12	205	73.70	52	84.62	45.20
13	979	73.95	⇐	13	19	81.60	42	80.95	59.10
14	410	71.95	⇐	14	31	78.90	42	78.57	54.20
15	285	71.23	⇐	15	57	84.20	42	85.71	50.80
16	157	75.16	⇐	16	188	80.30	58	79.31	54.50
17	896	76.00	⇐	17	31	76.30	44	86.36	66.10
18	683	74.96	⇐	18	334	75.00	46	84.78	49.90
19	204	72.06	⇐	19	53	81.60	40	85.00	56.60
20	678	74.63	⇐	20	141	78.90	42	85.71	60.50
21	1791	71.41	⇐	21	-5[a]	60.00	48	81.25	58.70
22	2058	67.69	⇐	22	42	85.50	54	87.04	56.60
23	1171	73.10	⇐	23	123	77.60	48	81.25	44.60
24	1759	69.93	⇐	24	-20[a]	73.70	40	82.50	58.00
25	2042	65.82	⇐	25	29	78.90	40	85.00	53.70
26	555	73.69	⇐	26	30	82.90	42	88.10	67.70
27	1947	65.59	⇐	27	23	85.50	40	87.50	67.20
28	1649	64.46	⇐	28	3	66.67	62	85.48	42.80
29	2489	67.34	⇐	29	-25[a]	76.00	40	77.50	41.20
30	1256	69.02							
Ø	1303.90	71.38		Ø	76.03	78.59	47.48	83.66	55.46
∑	39117			∑	2205		1384		

[a]negative numbers represent the length of an overlap, rather than a gap, of two adjacent contigs which could not be joined because of low overlapping quality and/or coverage.

[b]where 100% corresponds to a perfect hairpin (stem solely consisting of G-C pairs and the loop consisting of 3 unpaired bases).

to an uncovered region which would result in possible gap formation. Nevertheless, a GC-content above 75% was shown to be one of possibly multiple explanations for sequences not occurring in the sequencer's data output.

3.1.3. The gaps in the acarbose gene cluster are characterized by secondary structure formation

Although the GC-content shows a strong negative correlation with respect to the coverage, some gaps are also formed in regions with moderately high GC-content ($\leq$ 75%) whereas other regions with even higher GC-content do not decrease the coverage enough to cause contig breaks or gaps, respectively (**Fig. 3.2B&D**). This observation demands another parameter to explain all gaps in the cluster sequence. One reasonable possibility was derived from the observation that 12 of the 29 gaps were found to be completely located within intergenic regions (**Fig. 3.1**). In combination with the accompanying spikes in GC-content, these regions may well represent rho-independent intrinsic terminators, composed of strong secondary structures (stem-loops) which are formed by self-annealing of the single stranded DNA sequences. Using the `TransTermHP` [Kingsford *et al.*, 2007b] software, it was previously shown that it is possible to predict rho-independent transcription terminators composed of the typical palindromic region followed by a trail of thymidine residues [Banerjee *et al.*, 2006]. The `TransTermHP` search on the acarbose cluster reference sequence revealed one such terminator with high confidence exactly at the positions of the intergenic gap #26 (**Fig. 3.1**). Furthermore, it has recently been shown that the trail of thymidine residues is not necessarily required for the correct function of a terminator in certain organisms. In particular it was shown that a positive correlation exists between the GC-content of an organism and its prevalence for terminators without thymidine residues following the stem-loop structure [Mitra *et al.*, 2009, Unniraman *et al.*, 2001]. To account also for this kind of atypical terminators, the `GeSTer` prediction software [Mitra *et al.*, 2009] was applied without restriction to intergenic regions. Surprisingly, all terminator-like structures except those of the gaps #7, #12, #18, and #29 (**Fig. 3.4**) were identified by this approach although most of the structures are unlikely to function as transcriptional terminators due to their intragenic location (**Fig. 3.1**). As expected, most of the identified terminator-like structures are located within the gap regions or at least overlap them and thus validate the findings (**Fig. 3.5**).

In vivo, these terminators cause the RNA polymerase to dissociate from the template and thus terminate transcription of the molecule [Naville & Gautheret, 2009]. Similar effects have previously been reported during the amplification of DNA sequences containing strong secondary structures using standard PCR [McDowell *et al.*, 1998, Viswanathan *et al.*, 1999]. Furthermore, it was shown that secondary structures can increase the error frequency of DNA polymerases and therefore impact the amplification ability of the template significantly [Loewen & Switala, 1995]. According to these findings, it seems consequential that exactly the gap regions, containing terminator-like structures, could not be amplified in the emPCR step of the

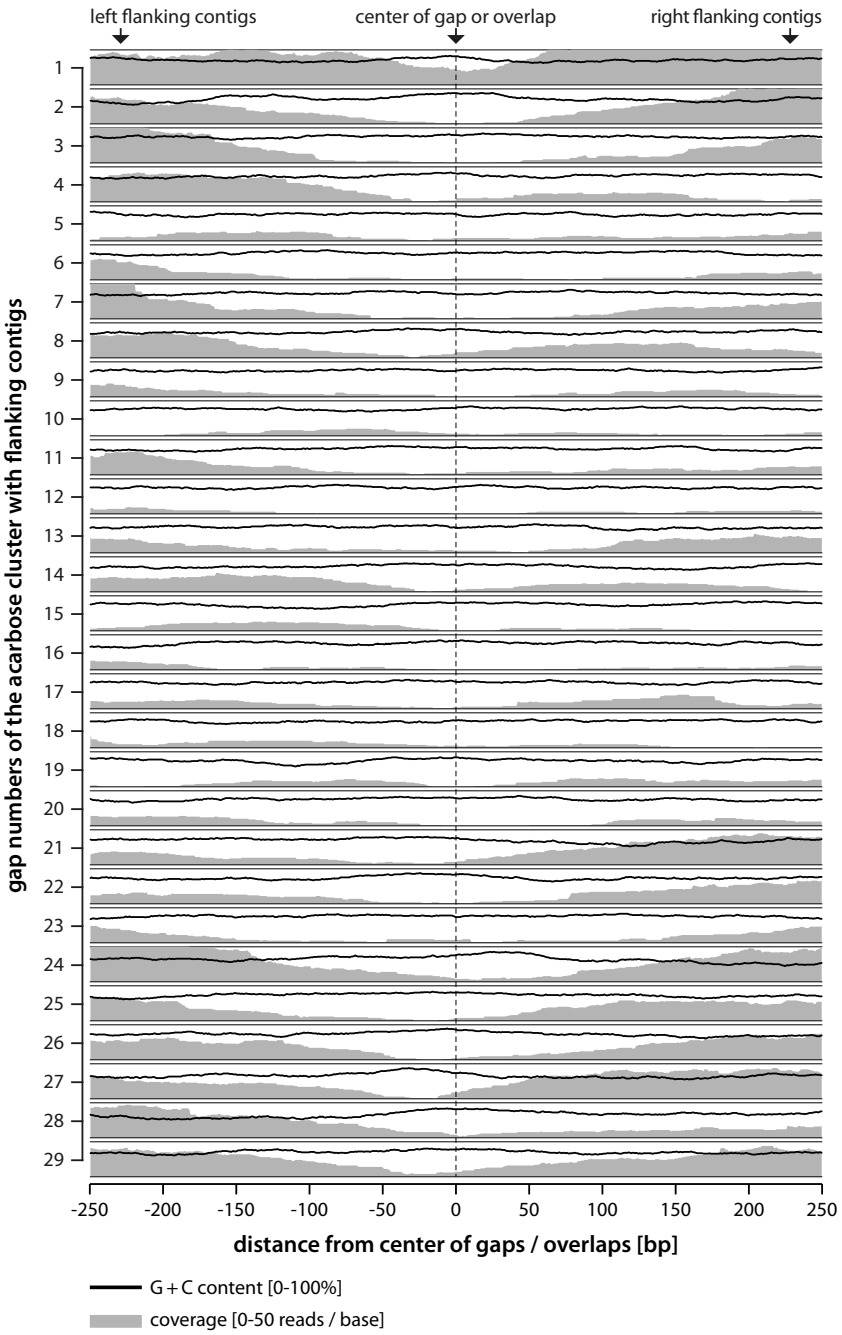

Figure 3.3.: Read coverage in the gaps of the acarbose gene cluster of *Actinoplanes* sp. SE50/110. Depicted are the 29 centered gaps as well as their flanking contigs (gray) together with the GC-content (black line) of the region.

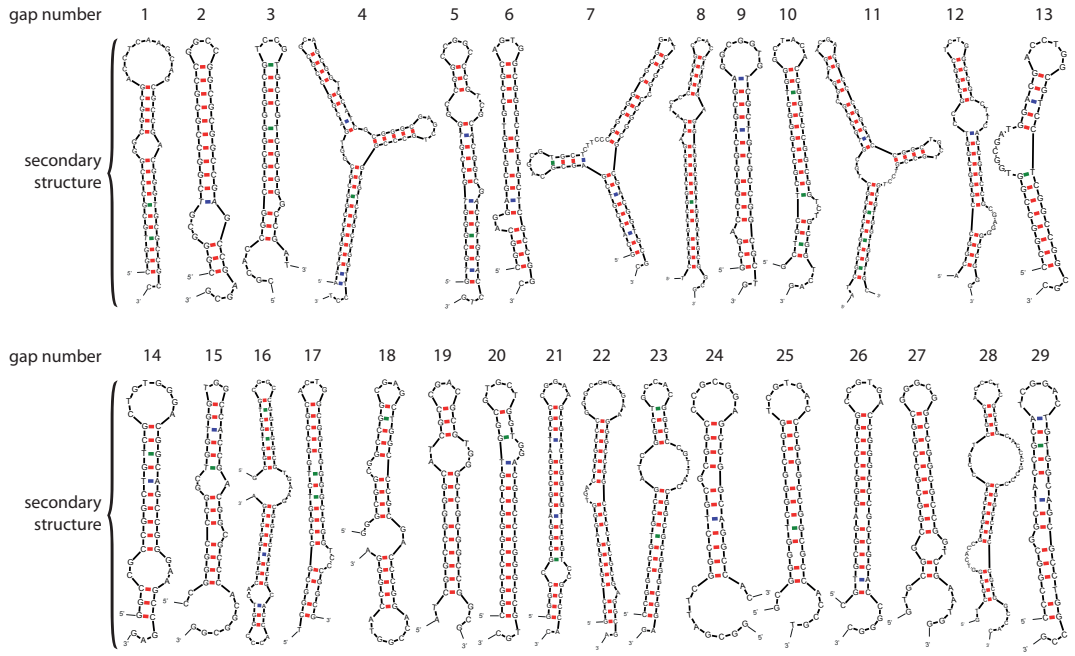

Figure 3.4.: Shapes of secondary structures found in gaps between adjacent contigs composing the acarbose biosynthetic gene cluster of *Actinoplanes* sp. SE50/110. These DNA sequences form strong hairpins by self-annealing which is supported by high-GC sequence content. The images were created by applying the `DNA mfold` software [Zuker, 2003] on 400 bases of flanking sequence of the center of gaps or overlaps between adjacent contigs. Subsequent reduction to the core structure resulted in the depicted secondary structures.

454 sequencing protocol, which would well explain their absence in the sequencer's data output.

Following these considerations, the ability of the remaining sequences to form stable self-annealing structures has been calculated by means of the `DNA mfold` software [Zuker, 2003]. The software calculates Gibbs free energy (ΔG) of a simulated self-annealing of the 76 base long sequence chunks, which is a measure for the ability of the sequence to form a stable secondary structure such as a stem-loop (**Fig. 3.4**). Surprisingly, numerous other regions with high self-annealing potential were discovered with this method (**Fig. 3.2C**). Similar to the negative correlation of the GC-content with the read coverage, the self-annealing and coverage also displays a high correlation coefficient of -0.47, demonstrating a decreasing coverage with rising ability for self-annealing and vice versa. Consequently, the GC-content and self-annealing also correlate strongly with a coefficient of 0.60, accounting for the stronger bond energy of G-C pairs as opposed to A-T base pairings during annealing.

Similar findings were previously reported for the formation of stable secondary structures like hairpins during the polymerase chain reaction which hamper the amplification of these sequences [Frey *et al.*, 2008]. To overcome this limitation and

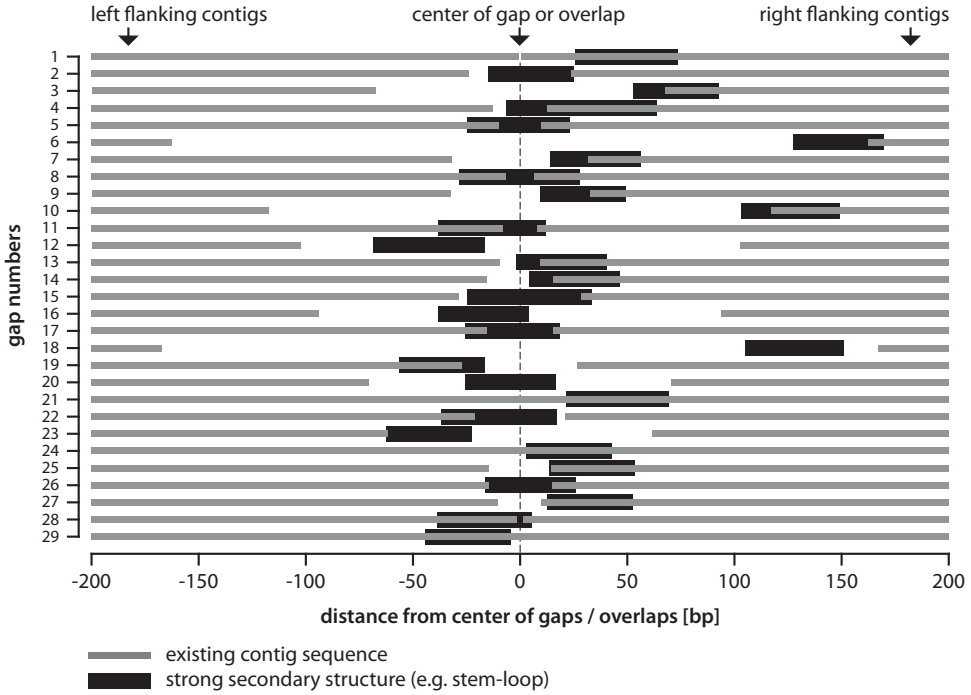

Figure 3.5.: The 29 gaps of the *Actinoplanes* sp. SE50/110 acarbose gene cluster and the positions of strong secondary structures (black) in relation to the adjacent contigs (gray). The depiction is centered to the middle of the gap or overlap (numbers #21, #24, and #29) and shows 200 bases of flanking region in each direction.

improve amplification of high GC-content DNA, several chemical PCR additives like deoxyinosine [Turner & Jenkins, 1995], trehalose [Spiess *et al.*, 2004] or betaine [Henke *et al.*, 1997, Weissensteiner & Lanchbury, 1996], and modified PCR protocols [Frey *et al.*, 2008, Sahdev *et al.*, 2007] were successfully applied.

Although the sequencing protocols and devices have changed tremendously between classical capillary and high-throughput sequencing techniques, all but the Helicos second generation high-throughput sequencing techniques routinely employ an initial amplification step. Since the basic PCR principle did also not change, it is consequential that these applications suffer from the same deficiencies in terms of high-GC sequence amplification bias as, up to now, no additive to prevent formation of secondary structures was supplemented to high-throughput amplification techniques such as the emulsion PCR of the GS FLX platform. Therefore it was concluded that strong secondary structures paired with high GC-content (**Fig. 3.2**) were formed during the emPCR of the 454 amplification step, causing the coverage to drop and violate the assembly conditions (40 bases overlap with 90% identity) under which reads are incorporated into existing contigs (see 'Genome Sequencer FLX System Software Manual', version 2.0).

3.1.4. Adapted sequencing conditions solved the high-GC problem

The identification of possible reasons for the formation of the large number of contigs led to a new sequencing attempt with two major differences. In order to inhibit the proposed self-annealing of DNA sequences, a new additive, called *emPCR Additive*, with inhibitory effects on the self-annealing ability of single stranded DNA molecules has been kindly provided by 454 Life Sciences. The additive has been substituted for the equal amount (1,500 µL) of H_2O in the emPCR chemistry. In addition, the WGS Titanium sequencing chemistry was used which featured an increased read length of ∼540 bases as opposed to ∼250 bases in the PE library (sequenced with the standard sequencing chemistry). With increased read length, small repetitive regions can be bridged more efficiently, guiding the assembly process towards fewer but larger contigs which were unconnected in the former PE library sequencing run with standard sequencing chemistry.

The WGS sequencing run was carried out on the same device, a Genome Sequencer FLX with Titanium sequencing chemistry. It yielded about 200 million bases in ∼480,000 reads (**Tab. 3.1**) with an inferred read-error of ∼0.5% (**Tab. 3.2**).

Strikingly, the results of the improved sequencing run exceeded all expectations by reducing the number of contigs from 7,973 to only 571 by ∼93% (**Tab. 3.3**). Furthermore, the average GC-content of the genome has increased by 0.87% to 71.31% whereas the genome length also increased significantly by ∼9% from 8.33 MB to 9.09 MB in comparison to the first sequencing. This provides strong evidence that the additionally sequenced DNA is very rich in GC-bases, often concentrated in sequences forming secondary structures as proposed above. Most notably, all gaps of the acarbose cluster reference sequence could be eliminated, resulting in a single contig (**Fig. 3.6**) despite the lower average genome coverage in comparison to the PE runs before, namely 21.25 fold opposed to 24.58 fold.

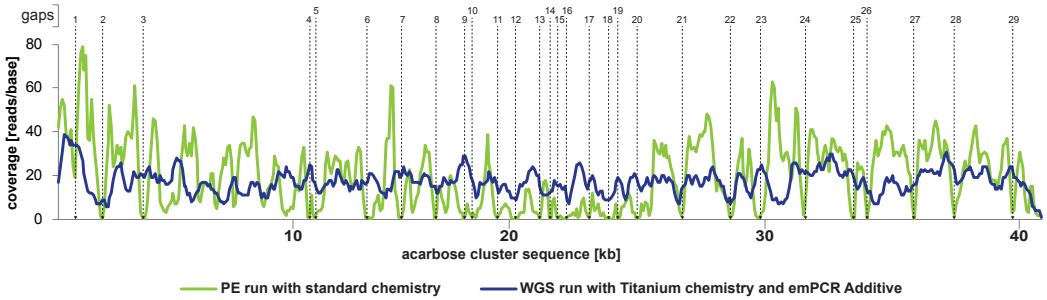

Figure 3.6.: Coverage chart of the 41 kb long acarbose biosynthetic gene cluster of *Actinoplanes* sp. SE50/110 for both sequencing runs. The coverage of the first paired end (PE) run with standard sequencing chemistry (green) is compared to the whole genome shotgun (WGS) run with Titanium sequencing chemistry and emPCR Additive (blue). The gap positions of the PE run are marked by vertical dashed lines.

Another positive side effect was the improved uniform distribution of reads across the cluster compared to the previous sequencing runs with the PE library and stan-

dard sequencing chemistry which may be especially interesting for quantitative RNA sequencing projects with high-GC genomes involved. Finally, the previously strong correlation between GC-content as well as the hybridization with the coverage has vanished, being as low as -0.09 and -0.11, respectively. To that effect, the read-coverage of both sequencing approaches correlated by 0.20 in a positive manner.

To elucidate the chemical nature of the applied emPCR Additive, the provided sample was analyzed via GC-MS and found to consist of trehalose. Previous experiments with trehalose as a PCR supplement reported an optimal concentration of 0.2 mol/L [Spiess *et al.*, 2004].

3.2. The complete genome sequence of *Actinoplanes* sp. SE50/110

3.2.1. Assembly of the *Actinoplanes* sp. SE50/110 draft genome sequence

The intermediate draft genome sequence was constructed by a combined assembly of all reads from both PE runs with standard chemistry and the reads from the WGS Titanium run. Although this assembly resulted in slightly more contigs than the WGS assembly, the overall quality increased significantly, which is best reflected by the reduced number of large contigs in conjunction with an increase in size of these contigs (**Tab. 3.5**). Put simply, the large contigs grew larger and were joined, whereas some new small contigs appeared. Likewise, the size of the draft genome increased to 9.15 Mb and the number of scaffolds dropped from 307 (PE runs) to only eleven in the combined assembly. Of these, three contained only a single contig, leaving only eight true scaffolds for further analysis.

Table 3.5.: Assembly results of combined PE and WGS sequencing runs for *Actinoplanes* sp. SE50/110.

Sequencing property	Results from combined PE and WGS assembly
No. of reads	1,968,468
Percent of aligned reads	99.65
Percent of aligned bases	98.72
No. of all contigs	600
No. of bases in all contigs	9,153,529
No. of large contigs ($\geq$ 500b)	476
No. of bases in large contigs	9,122,632
Percent of genome GC-content	71.27
Average genome coverage	43.88

The contigs of the draft genome were analyzed for over- or underrepresentation in read coverage by means of a scatter plot to identify repeats, putative plasmids or contaminations (**Fig. 3.7**). While most of the large contigs show an average coverage with reads (43.88 reads/base), several contigs were found to be clearly overrepresented

and are of special interest as discussed later. However, the majority of the unusually high and low covered contigs are of very short length, representing short repetitive elements (overrepresented) and contigs containing only few reads of low quality (underrepresented). These findings indicate clean sequencing runs without contaminations.

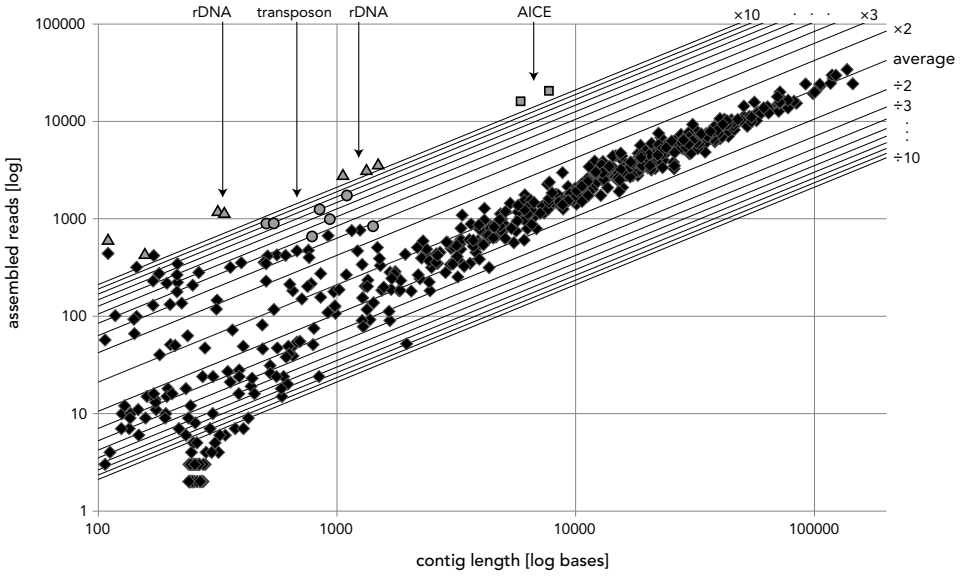

Figure 3.7.: Scatter plot of 600 *Actinoplanes* sp. SE50/110 contigs resulting from automatic combined assembly of the paired end and whole genome shotgun pyrosequencing runs. The average number of reads per base is 43.88 and is depicted in the plot by the central diagonal line marked with 'average'. Additional lines indicate the factor of over- and underrepresentation of reads per base up to a factor of 10 and 1/10 fold, respectively. The axes represent logarithmic scales. Large and highly overrepresented contigs are highlighted by special symbols. Each contig is represented by one of the following symbols: diamond, regular contig; square, contig related to an actinomycete integrative and conjugative element (AICE); triangle, contig related to ribosomal operon (*rrn*); circle, related to transposons

3.2.2. Finishing of the draft genome sequence

Based on PE information, eight scaffolds were constructed using 421 contigs with an estimated total length of 9,189,316 bases (**Fig. 3.8A**). These PE scaffolds were used to successfully map terminal insert sequences of 609 randomly selected fosmid clones from a previously constructed fosmid library with an insert size of ∼37 kb. The mapping results validated the PE scaffold assemblies and allowed the further assembly of the original eight paired end scaffolds into three PE/fosmid (PE/FO) scaffolds due to bridging fosmid reads (**Fig. 3.8B**).

Gap closure between the remaining contigs was carried out by fosmid walking (746 reads) and genomic PCR technology (236 reads) in cases where no fosmid was spanning

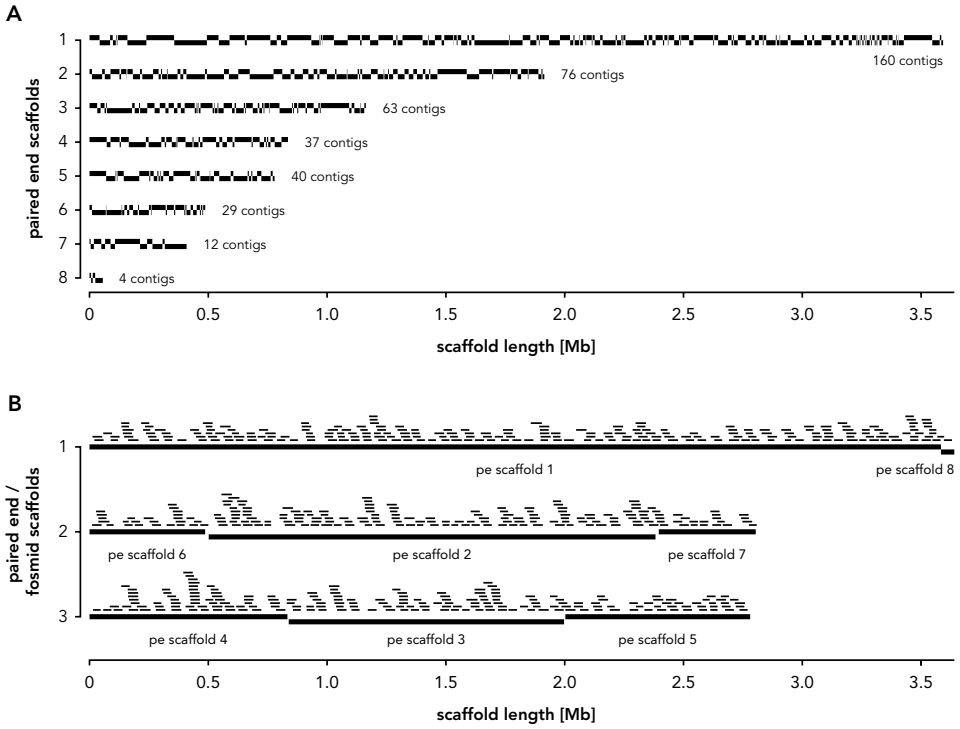

Figure 3.8.: Scaffolds of the *Actinoplanes* sp. SE50/110 genome. (**A**) The eight paired end (PE) scaffolds resulting from `Newbler` assembly of all paired end and whole genome shotgun reads are shown. Every second contig is visualized in a slightly displaced manner to show contig boundaries. (**B**) The three scaffolds resulting from terminal insert sequencing of fosmid (FO) clones and subsequent mapping on the PE scaffolds are presented. All overlapping sequences of the 609 mapped fosmid clones are shown on top of the PE/FO scaffolds.

the target region. Genomic PCR technology was also used to determine the order and orientation of the remaining three PE/FO scaffolds. The finishing procedure was manually performed using the `Consed` software [Gordon *et al.*, 1998] and resulted in the final assembly of a complete single circular chromosome of 9,239,851 bp with an average GC-content of 71.36% (**Fig. 3.9**). According to genome project standards [Chain *et al.*, 2009], the finished *Actinoplanes* sp. SE50/110 genome meets the gold standard criteria for high quality next generation sequencing projects. The general properties of the finished genome are summarized in **Table 3.6**.

3.2.3. Annotation of the complete genome sequence

Utilizing the prokaryotic gene finders `Prodigal` [Hyatt *et al.*, 2010] and `Gismo` [Krause *et al.*, 2007] in conjunction with the `GenDB` annotation pipeline [Meyer *et al.*, 2003], a total of 8,270 CDSs were determined on the *Actinoplanes* sp. SE50/110 genome (**Fig. 3.9**). These include 4,999 genes (60.5%) with an associated functional

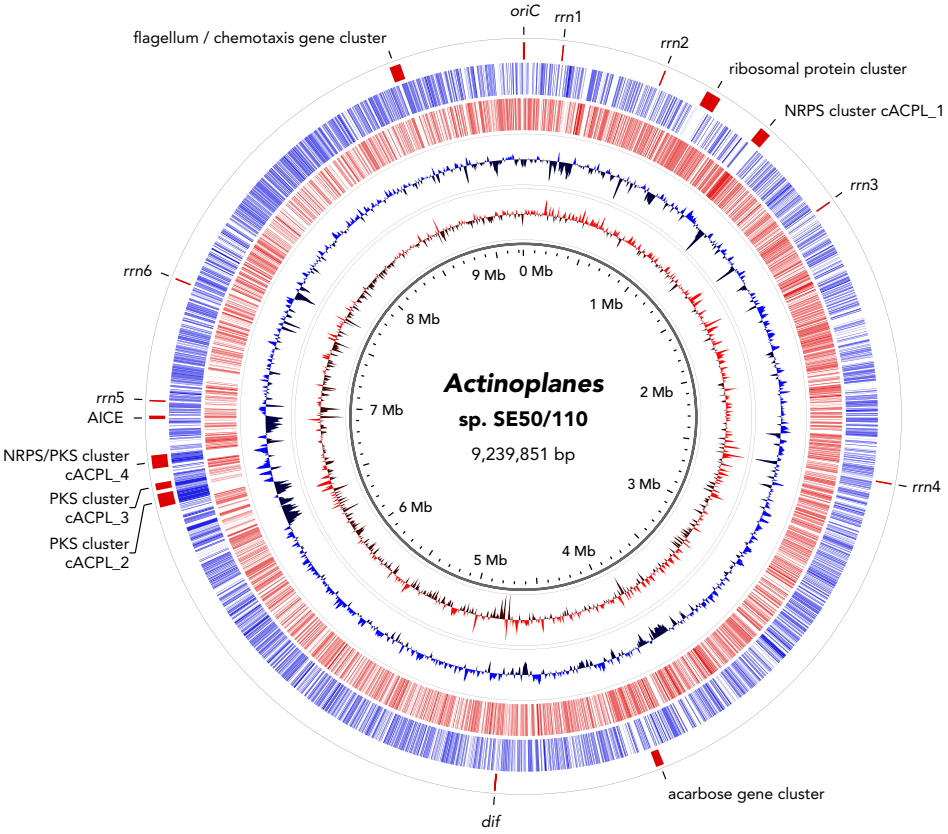

Figure 3.9.: Plot of the complete *Actinoplanes* sp. SE50/110 genome. The genome consists of 9,239,851 base pairs and 8,270 predicted coding sequences. The circles represent from the inside: 1, scale in million base pairs; 2, GC-skew; 3, GC-content; 4, genes in backward direction; 5, genes in forward direction; 6, gene clusters and other sites of special interest. Abbrevations were used as follows: *oriC* origin of replication, *dif* chromosomal terminus region, *rrn* ribosomal operon, NRPS non-ribosomal peptide synthetase, PKS polyketide synthase, AICE actinomycete integrative and conjugative element, cACPL cluster of *Actinoplanes*.

Table 3.6.: Features of the complete *Actinoplanes* sp. SE50/110 genome.

Feature	Chromosome
Total size (bp)	9,239,851
GC-content (%)	71.32
No. of protein-coding sequences	8,270
No. of orphans	973
Coding density (%)	89.31
Average gene length (bp)	985
No. of rRNAs	6 × 16S-23S-5S
No. of tRNAs	98

COG category [Tatusov *et al.*, 2001], 2,202 genes (26.6%) with a fully qualified EC-number [NC-ICBMB & Webb, 1992] and 973 orphan genes (11.8%) with neither annotation nor any similar sequence in public databases using BLASTP search with an e-value cutoff of 0.1. In total, the coding density of the genome amounts to 90.11% with a significant difference of 4% in GC-content between non-coding (67.74%) and coding (71.78%) regions.

Furthermore, 97 standard tRNA genes were determined by the `tRNAscan-SE` software [Lowe & Eddy, 1997] as well as one non-standard tRNA as described later in **Section 3.4.2**. **Figure 3.10** shows the absolute and relative gene counts of the twenty standard amino acids in relation to the occurrences of the corresponding amino acids derived from an analysis of all CDS of *Actinoplanes* sp. SE50/110. With the exception of alanine, which is the most often encoded amino acid in *Actinoplanes* sp. SE50/110, the ratios between tRNA genes and encoded amino acids correlates quite good with a correlation coefficient of 0.79.

The complete annotated genome sequence was deposited at the National Center for Biotechnology Information (NCBI) [GenBank:CP003170].

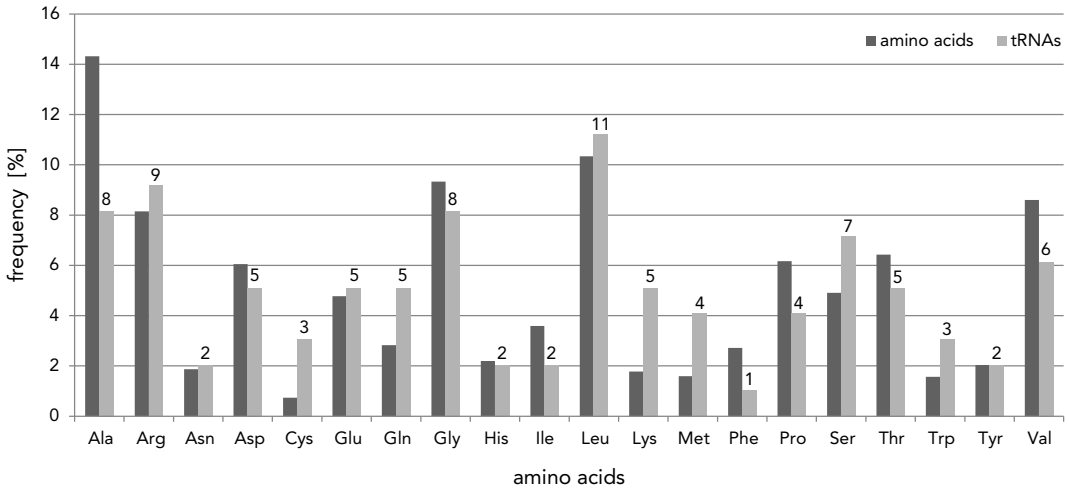

Figure 3.10.: Comparison of occurences of tRNAs and the corresponding amino acids encoded in all CDS of the *Actinoplanes* sp. SE50/110 genome. Values are depicted in percent on the vertical axis and given in absolute numbers for the tRNA genes at their corresponding bars.

3.3. Discoveries of the *Actinoplanes* sp. SE50/110 genome

3.3.1. General features of the *Actinoplanes* sp. SE50/110 genome

The origin of replication (*oriC*) was identified as a 1266 bp intergenic region between the two genes *dnaA* and *dnaN*, coding for the bacterial chromosome replication initiator protein and the β-sliding clamp of the DNA polymerase III, respectively. The *oriC*

harbors 24 occurrences of the conserved DnaA box [TT(G/A)TCCACa], showing remarkable similarity to the *oriC* of *Streptomyces coelicolor* [Zawilak-Pawlik *et al.*, 2005]. Almost directly opposite of the *oriC*, a putative *dif* site was found. Its 28 bp sequence 5'-CAGGTCGATAATGTATATTATGTCAACT-3' is in good accordance with actinobacterial *dif* sites and shows highest similarity (only 4 mismatches) to that of *Frankia alni* [Hendrickson & Lawrence, 2007]. In addition to the identified *oriC* and *dif* sites, the calculated G/C skew [(G-C)/(G+C)] suggests two replichores composing the circular *Actinoplanes* sp. SE50/110 genome (**Fig. 3.9**).

In accordance with previous findings [Mehling *et al.*, 1995b], six ribosomal RNA (*rrn*) operons were identified on the genome in the typical 16S-23S-5S order. The six individual *rrn* operons were previously assembled into one operon ranging across five contigs with a more than ten-fold overrepresentation (**Fig. 3.7**). The discrepancy between the six actual *rrn* operons and a more than ten-fold overrepresentation of their representative contigs may be explained by the operon's remarkably low GC-content of 57.20% in comparison to the genome average of 71.36%, which is obviously typical for many actinomycetes [Mehling *et al.*, 1995b]. The low GC-content in this area may have introduced an amplification bias in favor of the *rrn* operon during the library preparation and, thus, resulted in an overrepresentation of reads for this genomic region. To account for SNPs and variable regions between ribosomal genes, all six *rrn* operons were individually re-sequenced by fosmid walking. The *rrn* operons are located on the leading strands, four on the right and two on the left replichore. Interestingly, they reside in the upper half of the genome, together with a ∼40 kb gene cluster hosting more than 30 ribosomal proteins (**Fig. 3.9**). Other large overrepresented contigs were identified as transposase genes or transposon related elements (**Fig. 3.7**).

Approximately 500 kb upstream of the *oriC* site, a flagellum gene cluster was found. Its expression in spores is one of the characteristics discriminating the genus *Actinoplanes* from other related species [Couch, 1950, Parenti & Coronelli, 1979]. The cluster consists of ∼50 genes spanning 45 kb. Besides flagella associated proteins, the cluster also contains genes coding for chemotaxis related proteins.

The codon usage of *Actinoplanes* sp. SE50/110 reveals a strong prevalence for codons ending an a guanine or cytosine base (**Fig. 3.11**). On average, the GC-content of the codons are 71.8% on the first, 51.7% on the second, and 91.8% on the third letter of a codon. Furthermore, a more detailed analysis of the start and stop codons showed that more than 65% of all CDSs start with an ATG codon, whereas TGA dominates among the stop codons and is found in more than 75% of all CDSs in *Actinoplanes* sp. SE50/110 (**Fig. A.1**).

Bioinformatic classification of 4,999 CDSs with an annotated COG-category revealed a strong emphasis (47%) on enzymes related to metabolism (**Fig. 3.12**). In particular, *Actinoplanes* sp. SE50/110 features an emphasis on amino acid (10%) and carbohydrate metabolism (11%) which is in good accordance with the identification of at least 29 ABC-like carbon substrate importer complexes. Furthermore, 16% of the COG-classified CDSs code for proteins involved in transcriptional processes which suggests a high level of regulation in the expression of various biosynthetic pathways.

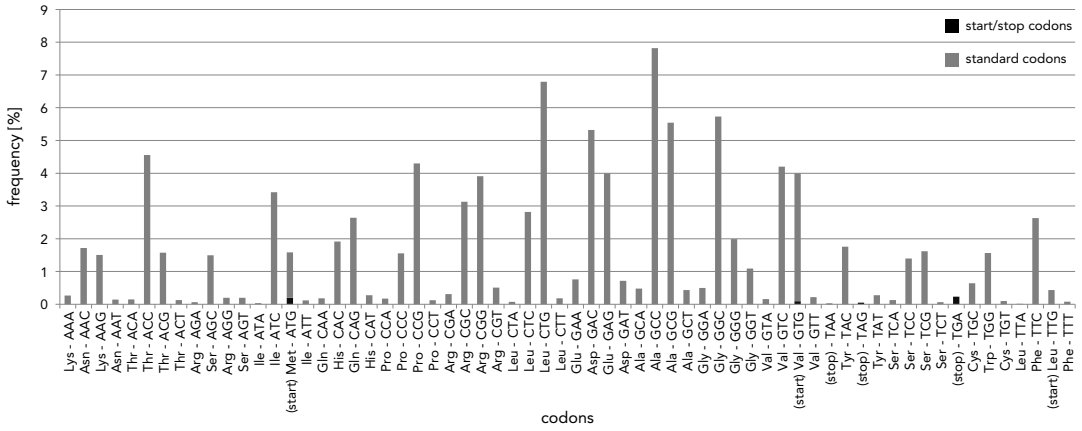

Figure 3.11.: Codon usage of *Actinoplanes* sp. SE50/110 based on all 8.270 CDSs (2,728,490 codons).

This is especially relevant for the ongoing search for a regulatory element or -network controlling the expression of the acarbose biosynthetic gene cluster. Interestingly, the great proportion of transcriptional regulators is accompanied by a similar high percentage of proteins involved in signal transduction mechanisms (12%) which suggests a close connection between extracellular nutrient sensing and transcriptional regulation of uptake systems and degradation pathways. In contrast to *Actinoplanes* sp. SE50/110, the results of an analogous analysis of 4,431 annotated CDSs from *Streptomyces coelicolor* revealed only 7% of the encoded proteins being involved in signal transduction mechanisms whereas the amount of proteins involved in transcriptional processes is highly similar (16%). Overall, *S. coelicolor* hosts even more genes coding for enzymes related to metabolism (55%), whereas the genome of *Actinoplanes* sp. SE50/110 reveals a striking focus (27%) on cellular processes and signaling when compared to *S. coelicolor* (20%). Besides these findings, only the genes for carbohydrate transport and -metabolism shows another notable difference of more than 1% between *S. coelicolor* (13%) and *Actinoplanes* sp. SE50/110 (11%). Interestingly, 4% of the *Actinoplanes* sp. SE50/110 CDS were found to be involved in secondary metabolite biosynthesis (*S. coelicolor* 5%). Taken together, these considerations lead to a new perception of the capabilities *Actinoplanes* sp. SE50/110 might offer. Rather than being the producer of acarbose, *Actinoplanes* sp. SE50/110 features a large amount of genes that could encode secondary metabolite biosynthesis pathways comparable to that found in well-known producers like *Streptomyces coelicolor* [Bentley *et al.*, 2002] or *Salinispora tropica* [Udwary *et al.*, 2007]. Furthermore and in contrast to *S. coelicolor*, the genome of *Actinoplanes* sp. SE50/110 hosts significantly more genes for signal transduction proteins. This might be one key to induce the expression of acarbose and novel secondary metabolite gene clusters by appropriately composed cultivation media following the one strain, many compounds (OSMAC) approach [Höfs *et al.*, 2000]. These considerations are in good accordance with empirical knowledge

gathered through long lasting media optimizations [BAYER HEALTHCARE, PERSONAL COMMUNICATION].

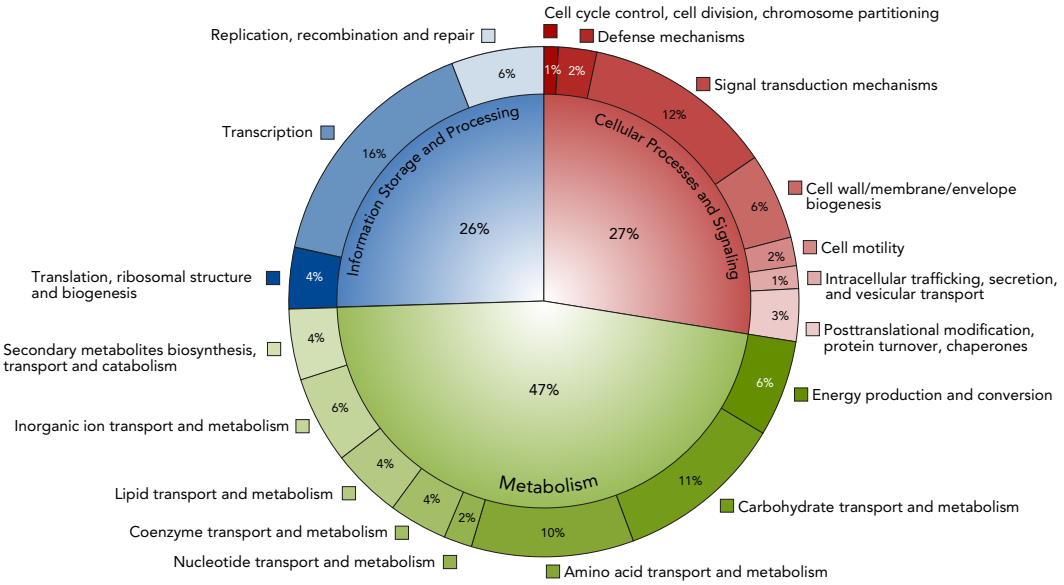

Figure 3.12.: Functional classifications of the *Actinoplanes* sp. SE50/110 CDSs. The diagram represents the CDSs that were categorized according their cluster of COG number [TATUSOV *et al.*, 1997, TATUSOV *et al.*, 2001]. All depicted percentages refer to the distribution of 4999 annotated CDSs (100%) across all COG categories to which at least 10 CDSs were found. Sequences with an unknown or poorly characterized function were excluded from the analysis. The outer ring contains specialized subclasses of the three main functional categories 'Cellular Processes and Signaling', 'Metabolism', and 'Information Storage and Processing', located at the center.

3.3.2. Phylogenetic analysis of the *Actinoplanes* sp. SE50/110 16S rDNA reveals highest similarity to *Actinoplanes utahensis*

An unsupervised nucleotide `BLAST` [ALTSCHUL *et al.*, 1990] run of the 1509 bp long DNA sequence of the 16S rRNA gene from *Actinoplanes* sp. SE50/110 against the public non-redundant database (NCBI nr/nt) revealed high similarities to numerous species of the genera *Actinoplanes*, *Micromonospora* and *Salinispora*. Within the best 100 matches, the maximal DNA sequence identity was in the range of 100-96%. The coverage of the query sequence varied within this cohort between 100-97%. The hits with the highest similarity, based on the number of sequence substitutions were *A. utahensis* IMSNU 20044^{T} (17 substitutions, 3 gaps) and *A. utahensis* IFO 13244^{T} (16 substitutions, 3 gaps) which both retrace to the type strain (^{T}) *A. utahensis* ATCC 14539^{T} described first by John Nathaniel Couch in 1963 [COUCH, 1963]. The third hit to *A. palleronii* IMSNU 2044^{T} differs from *Actinoplanes* sp. SE50/110 by 24 substitutions and 5 gaps.

Based on the multiple sequence alignment of the best 100 BLAST hits, a phylogenetic tree was derived as described in **Section 2.5.3**. A detailed view on a subtree contains *Actinoplanes* sp. SE50/110 and 34 of the most closely related species (**Fig. 3.13**). This subtree displays the derived phylogenetic distances between the analyzed strains, represented by their distance on the x-axis. From this analysis, it is evident that, based on 16S rDNA comparison, *A. utahensis* is the nearest species to *Actinoplanes* sp. SE50/110 currently publicly known, followed by *A. palleronii* and *A. awajiensis* subsp. *mycoplanecinus*. A second analysis using the latest version of the ribosomal database project [Cole *et al.*, 2009] resulted in highly similar findings (data not shown). Interestingly, *A. utahensis* and *Actinoplanes* sp. SE50/110 form an encapsulated subcluster within the *Actinoplanes* genus although the different isolates originate from far distant locations on different continents (Salt Lake City, USA, North America and Ruiru, Kenya, Africa). In addition, it is noteworthy that *Actinoplanes* sp. SE50/110 was renamed several times and in the early 1990s this strain was also classified as *A. utahensis* [Mehling *et al.*, 1995b].

3.3.3. Comparative genome analysis indicates 50% singletons in the *Actinoplanes* sp. SE50/110 genome.

To date, seven full genome sequences belonging to the family Micromonosporaceae are publicly available. Using the comparative genomics tool EDGAR [Blom *et al.*, 2009], a gene based, full genome phylogenetic analysis of these strains revealed a comparable phylogeny as was derived for the 16S rDNA based method (**Fig. 3.14**). For comparison, some industrially used *Streptomyces* and *Frankia* strains were also included in the analysis. As expected, each genus forms its own cluster. Interestingly, the genera *Micromonospora*, *Verrucosispora* and *Salinispora* are more closely related to each other than to *Actinoplanes*, whereas *Streptomyces* and *Frankia* are clearly distinct from the whole Micromonosporaceae family. Based on this analysis, the marine sediment isolate *Verrucosispora maris* AB-18-032 is the closest sequenced species to *Actinoplanes* sp. SE50/110 currently publicly known with 2,683 orthologous genes, a GC-content of 70.9% and a genome size of 6.67 MB [Roh *et al.*, 2011]. Comparative BLAST analysis of conserved orthologous genes of all sequenced Micromonosporaceae strains revealed prevalence for being located in the upper half of the genome, near the origin of replication (data not shown). The core genome analysis revealed a total of 1,670 genes common to all seven *Micromonosporaceae* strains, whereas the pan genome consists of 18,189 genes calculated by the EDGAR software. Analysis of the singletons revealed 4,122 genes (49.8%) exclusively occurring in the *Actinoplanes* sp. SE50/110 genome, not present on the other six Micromonosporaceae strains.

3.3.4. The high quality genome sequence of *Actinoplanes* sp. SE50/110 corrects the previously sequenced acarbose cluster.

The first sequence fraction of the acarbose biosynthetic (*acb*) gene cluster was initially identified [Stratmann *et al.*, 1999] and successively expanded by classical Sanger se-

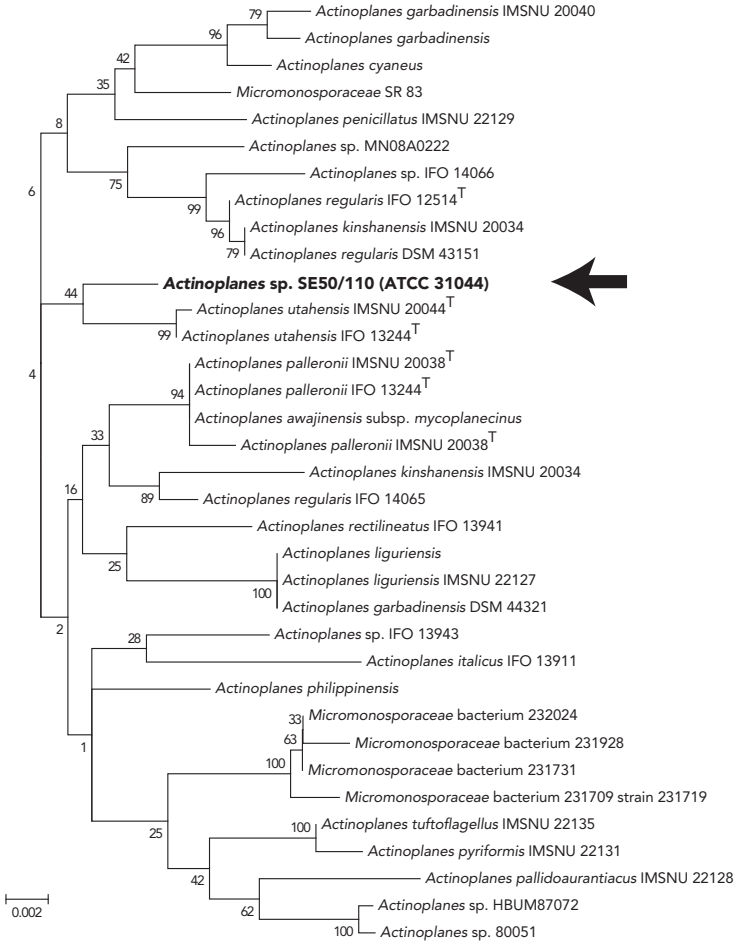

Figure 3.13.: Phylogenetic tree based on 16S rDNA from *Actinoplanes* sp. SE50/110 and the 34 most closely related species. Shown is an excerpt of a phylogenetic tree built from the 100 best nucleotide BLAST hits for the *Actinoplanes* sp. SE50/110 16S rDNA. The shown subtree contains the 34 hits most closely related to *Actinoplanes* sp. SE50/110 (black arrow) with their evolutionary distances. The numbers on the branches represent confidence values in percent from a phylogenetic bootstrap test (1000 replications). The evolutionary history was inferred using the Neighbor-Joining method [Saitou & Nei, 1987]. The bootstrap consensus tree inferred from 1000 replicates is taken to represent the evolutionary history of the taxa analyzed [Felsenstein, 1985]. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Jukes-Cantor method [Jukes & Cantor, 1969] and are in the units of the number of base substitutions per site. The analysis involved 100 nucleotide sequences of which 35 are shown. Codon positions included were 1st+2nd+3rd+Noncoding. All positions containing gaps and missing data were eliminated. There were a total of 1396 positions in the final dataset. Evolutionary analyses were conducted in MEGA5 [Tamura *et al.*, 2007]. The scale represents 0.002 nucleotide substitutions per nucleotide position.

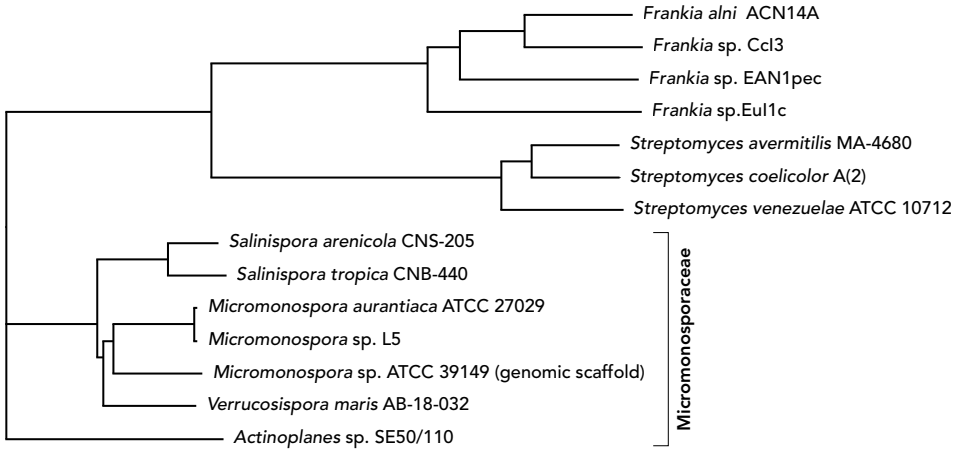

Figure 3.14.: Phylogenetic tree based on CDSs from *Actinoplanes* sp. SE50/110 and six species of the family Micromonosporaceae as well as *Streptomyces* and *Frankia* strains. The tree was constructed using the software tool EDGAR [Blom *et al.*, 2009] based on 605 core genome CDSs from the species occurring in the analysis. The comparison shows all seven strains of the taxonomic family Micromonosporaceae sequenced and publicly available to date in relation to other well studied bacteria.

quencing [Hemker *et al.*, 2001, Wehmeier & Piepersberg, 2004]. Until now, this sequence was the longest (41,323 bp [GenBank:Y18523.4]) and best studied contiguous DNA fragment available from *Actinoplanes* sp. SE50/110. However, with the complete, high quality genome at hand, a total of 61 inconsistent sites were identified in the existing acarbose gene cluster sequence (**Fig. 3.15**). Most notably, the deduced corrections affect the amino acid sequence of two genes, namely *acbC*, coding for the cytoplasmic 2-*epi*-5-*epi*-valiolone-synthase, and *acbE*, translating to a secreted long chain acarbose resistant α-amylase [Hemker *et al.*, 2001, Wehmeier & Piepersberg, 2004]. Because of two erroneous nucleotide insertions (c.1129_1130insG and c.1146_1147insC) in *acbC* (1197 bp), the resulting frameshift caused a premature stop codon to occur, shortening the actual gene sequence by 42 nucleotides. In contrast to *acbC*, the sequence differences in *acbE* (3102 bp) are manifold, including mismatches, insertions and deletions which lead to multiple temporary frameshifts and single amino acid substitutions in the middle part of the gene sequence ranging from nucleotide position 1102 to 2247. Even though these sequence corrections are important and improved the similarity of the α-amylase domain to its catalytic domain family, the overall annotated function of both gene products remains valid.

3.3.5. Several genes of the acarbose gene cluster are also found in other locations of the genome.

It is known that the copy number of genes can have a high impact on the efficiency of secondary metabolite production [Baltz, 1998, Baltz, 2001, Olano *et al.*, 2008,

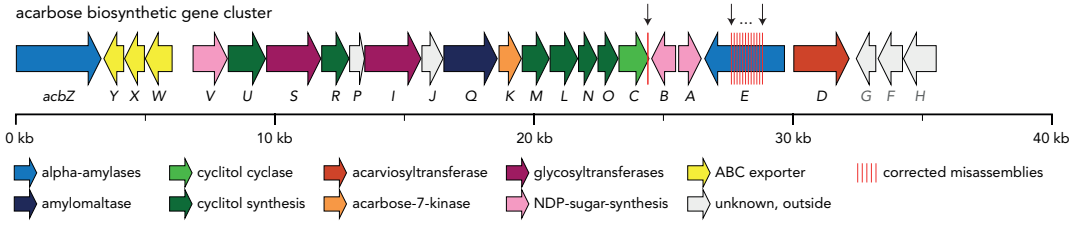

Figure 3.15.: The structure of the acarbose biosynthetic gene cluster from *Actinoplanes* sp. SE50/110. Based on the whole genome sequence, several nucleotide corrections were found with respect to the previously sequenced reference sequence of the acarbose gene cluster [GenBank:Y18523.4]. The corrected sites in *acbC* and *acbE* are marked by arrows and red dashes.

Baltz, 2011]. It is therefore worthwhile to study the genome wide occurrences of the genes encoded within the acarbose biosynthetic gene cluster, particularly with regard to import and export systems and the assessment of possible future knock-out experiments.

The results show that the *acb* gene cluster does not occur in more than one location within the *Actinoplanes* sp. SE50/110 genome. However, single genes and gene sets with equal functional annotation and amino acid sequence similarity to members of the *acb* cluster scattered throughout the genome were found by `BLASTP` analysis. Most notably, homologues to genes encoding the first, second and fourth step of the valienamine moiety synthesis of acarbose were found as a putative operon with moderate similarities of 52% (Acpl6250 to AcbC), 35% (Acpl6249 to AcbM) and 34% (Acpl6251 to AcbL). Furthermore, one homologue for each of the proteins AcbA (61% to Acpl3097) and AcbB (66% to Acpl3096) was identified. While the genes *acbA* and *acbB* are located adjacent to each other on the acarbose gene cluster (**Fig. 3.15**), they were also shown to catalyze the first two sequential reactions needed for the formation of dTDP-4-keto-6-deoxy-D-glucose, another essential intermediate in the acarbose biosynthesis [Stratmann *et al.*, 1999, Wehmeier, 2003]. It is therefore interesting to note that the identified homologues to *acbA* and *acbB* were also found adjacent to each other in the context of a putative dTDP-rhamnose synthesis cluster (*acpl3095-acpl3098*), which was shown to code for mandatory proteins RmlABCD involved in cell wall integrity and, thus, survival of *Mycobacterium smegmatis* [Li *et al.*, 2006]. Further genome analysis revealed two homologous operons to the acarbose exporter complex AcbWXY. Acpl3214-Acpl3216, showing 30-44% and Acpl5011-Acpl5013, indicating 28-49% sequence similarity. Both operons resemble the gene structure of *acbWXY* consisting of an ABC-type sugar transport ATP-binding protein and two ABC-type transport permease protein coding genes. In case of Acpl5011-Acpl5013, the CDS for a second ATP-binding protein, Acpl5010 overlaps the 5'-start of *acpl5011* by 46 bases and is therefore likely to belong to the operon as well. The similarities to other characterized ABC-type transporter complexes are too low to allow reliable conclusions about the substrate specificity. Acpl6399, a homologue with high sequence

similarities to the alpha amylases AcbZ (65%) and AcbE (63%), was found encoded within the maltose importer operon *malEFG*. For the remaining *acb* genes only weak (*acbV*, *acbR*, *acbP*, *acbJ*, *acbQ*, *acbK* and *acbN*) or no similarities (*acbU*, *acbS*, *acbI* and *acbO*) were found outside of the *acb* cluster by `BLASTP` searches using an e-value threshold of $1e^{-10}$.

3.3.6. Trehalose synthesis in *Actinoplanes* sp. SE50/110

Trehalose is a non-reducing disaccharide which is utilized in a wide variety of living organisms. Among other functions, it serves as an osmoprotector that protects cells from dehydration and can also be used as energy source [Avonce *et al.*, 2006]. Especially bacterial spores exhibit a high concentration of trehalose. In respect to the acarbose production, the relevance of trehalose is given by its substitution with the maltose residue of the acarbose molecule and hence, the formation of component C (see **Figure 1.2** and **Table 1.1**).

A genome wide search for genes encoding trehalose synthases revealed nine genes which are putatively involved in this reaction (**Tab. 3.7**). These belong to three of the six known pathways for trehalose synthesis [Avonce *et al.*, 2006]. The first pathway (TPS/TPP) involves two enzymes, trehalose 6-phosphate synthase (TPS), which catalyses the reaction UDP-glucose + glucose 6-phosphate $\mapsto$ trehalose 6-phosphate and trehalose 6-phosphate phosphatase (TPP), which catalyzes the reaction trehalose 6-phosphate $\mapsto$ trehalose. The *Actinoplanes* genome hosts four putative TPS encoding genes (*otsA*) and one TPP encoding gene (*otsB*). However, none of these are located within a common gene cluster. The second pathway consists of a single trehalose synthase (TS), which is capable of isomerizing maltose directly into trehalose. Two such encoding genes (*treS*) were found in the genome. The third pathway (TreY/TreZ) contains again two enzymes. The maltooligosyl-trehalose-synthase TreY, which converts maltooligosaccharides, starch or glycogen to maltooligosyltrehalose, and the maltooligosyl-trehalose trehalohydrolase TreZ, which catalyzes the further conversion to trehalose. One cluster containing both genes (*treY* and *treZ*) was found together with a third gene *treX*, encoding a glycogen debranching enzyme which is not directly involved in the trehalose synthesis.

3.3.7. The *Actinoplanes* sp. SE50/110 genome hosts an integrative and conjugative element

The actinomycete integrative and conjugative elements (AICEs) are a class of mobile genetic elements possessing a highly conserved structural organization with functional modules for excision/integration, replication, conjugative transfer and regulation [te Poele *et al.*, 2008]. Being able to replicate autonomously, they are also said to mediate the acquisition of additional modules, encoding functions such as resistance and metabolic traits, which confer a selective advantage to the host under certain environmental conditions [Burrus & Waldor, 2004]. Interestingly, a similar AICE, designated plasmid of *Actinoplanes* (pACPL), was identified in the complete genome

Table 3.7.: Trehalose synthases of *Actinoplanes* sp. SE50/110.

Pathway	Locus tag	Protein length	Gene symbol	Description
TPS/TPP	*acpl2177*	471	*otsA*	trehalose-6-phosphate synthase
TPS/TPP	*acpl1678*	472	*otsA*	trehalose-6-phosphate synthase
TPS/TPP	*acpl1307*	501	*otsA*	trehalose-6-phosphate synthase
TPS/TPP	*acpl3417*	481	*otsA*	trehalose-6-phosphate synthase
TPS/TPP	*acpl7709*	269	*otsB*	trehalose-6-phosphate phosphatase
TS	*acpl5330*	586	*treS*	trehalose synthase
TS	*acpl7518*	564	*treS*	trehalose synthase
TreY/TreZ	*acpl6623*	704	*treX*	glycogen debranching enzyme
TreY/TreZ	*acpl6624*	756	*treY*	maltooligosyltrehalose synthase
TreY/TreZ	*acpl6625*	578	*treZ*	maltooligosyltrehalose trehalohydrolase

sequence of *Actinoplanes* sp. SE50/110 (**Fig. 3.16**). Its size of 13.6 kb and the structural gene organization are in good accordance with other known AICEs of closely related species like *Micromonospora rosario*, *Salinispora tropica* or *Streptomyces coelicolor* (**Fig. 3.16**).

Most known AICEs subsist in their host genome by integration in the 3'-end of a tRNA gene by site-specific recombination between two short identical sequences (*att* identity segments) within the attachment sites located on the genome (*attB*) and the AICE (*attP*), respectively [te Poele *et al.*, 2008]. In pACPL, the *att* identity segments are 43 bp in size and *attB* overlaps the 3'-end of a proline tRNA gene. Moreover, the identity segment in *attP* is flanked by two 21 bp repeats containing two mismatches: `GTCACCCAGTTAGT(T/C)AC(C/T)CAG`. These exhibit high similarities to the arm-type sites identified in the AICE pSAM2 from *Strepomyces ambofaciens*. For pSAM2 it was shown that the integrase binds to these repeats and that they are essential for efficient recombination [Raynal *et al.*, 2002].

pACPL hosts 22 putative protein coding sequences (**Fig. 3.16**). The integrase, excisionase and replication genes *int*, *xis* and *repSA* are located directly downstream of *attP* and show high sequence similarity to numerous homologues from closely related species. The putative main transferase gene *tra* contains the sequence of a FtsK-SpoIIIE domain found in all pACPLs and *Streptomyces* transferase genes [te Poele *et al.*, 2008]. SpdA and SpdB show weak similarity to spread proteins from *Frankia* sp. CcI3 and *M. rosaria* where they are involved in the intramycelial spread of pACPLs [Kataoka *et al.*, 1994, Grohmann *et al.*, 2003]. The putative regulatory protein Pra was first described in pSAM2 as an pACPL replication activator [Sezonov *et al.*, 1995]. On pACPL, it exhibits high similarity to an uncharacterized homologue from *Micromonospora aurantica* ATCC 27029. A second regulatory gene reg shows high similarities to transcriptional regulators of various Streptomyces strains whereas the downstream gene *nud* exhibits 72% similarity to the amino acid sequence of a NUDIX

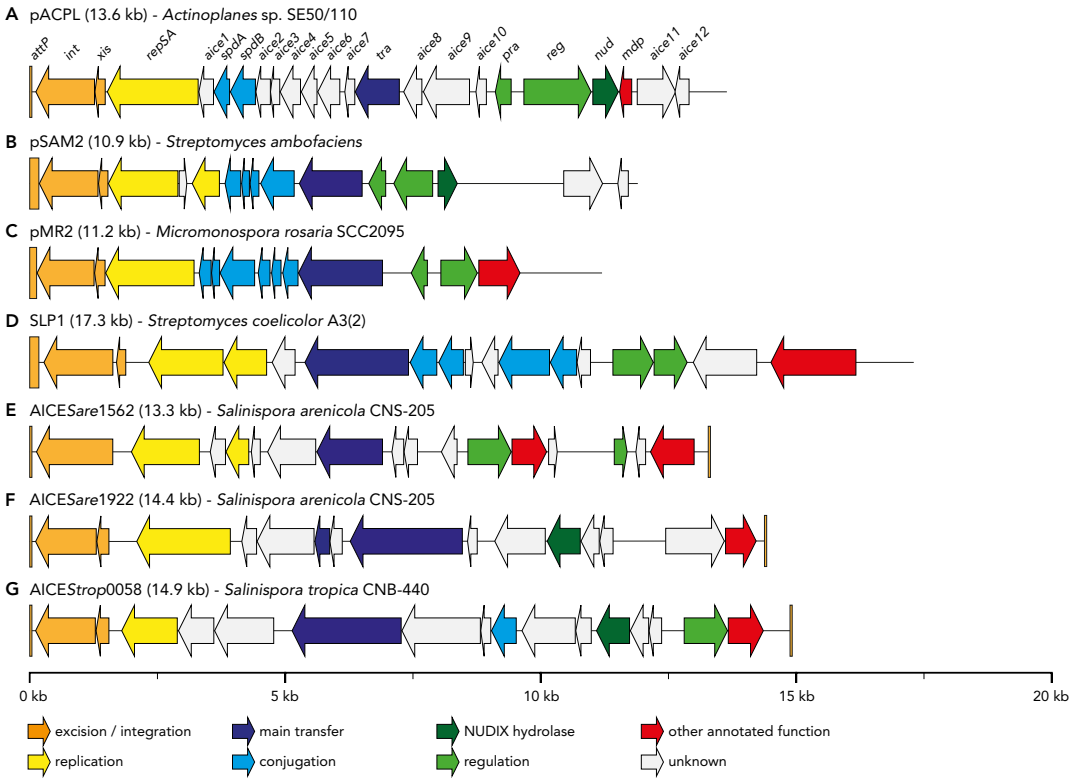

Figure 3.16.: Structural organization of the newly identified AICE pACPL from *Actinoplanes* sp. SE50/110 in comparison with other AICEs from closely related species. (**A**) pACPL (13.6 kb), the first AICE found in the Actinoplanes genus from *Actinoplanes* sp. SE50/110; (**B**) pSAM2 (10.9 kb) from *Streptomyces ambofaciens*; (**C**) pMR2 (11.2 kb) from *Micromonospora rosaria* SCC2095; (**D**) SLP1 (17.3 kb) from *Streptomyces coelicolor* A3(2); (**E**, **F**) AICESare1562 (13.3 kb) and AICESare1922 (14.4 kb) from *Salinispora arenicola* CNS-205; (**G**) AICEStrop0058 (14.9 kb) from *Salinispora tropica* CNB-440. Typical genes found in AICEs are colored: excision / integration (orange), replication (yellow), main transfer (dark blue), conjugation (blue), NUDIX hydrolase (dark green), regulation (green), other annotated function (red), unknown function (gray). B-G adapted from [te Poele *et al.*, 2008]

hydrolase from *Streptomyces* sp. AA4. In contrast, *mdp* codes for a metal dependent phosphohydrolase also found in various *Frankia* and *Streptomyces* strains.

Homologues to the remaining genes are poorly characterized and largely hypothetical in public databases although *aice4* is also found in various related species and shows, akin to *aice1*, *aice2*, *aice5*, *aice6*, and *aice9*, high similarity to homologues from *M. aurantiaca*. Interestingly, homologues to *aice1* and *aice2* were only found in *M. aurantiaca*, whereas *aice3*, *aice7*, *aice8*, *aice10*, *aice11*, and *aice12* seem to solely exist in *Actinoplanes* sp. SE50/110.

Based upon read-coverage observations of the AICE containing genomic region, an approximately twelve-fold overrepresentation of the AICE coding DNA sequences has been revealed (**Fig. 3.7**). As only one copy of the AICE was found to be integrated in the genome, it was concluded that on average about eleven copies of the element exist as circular, extrachromosomal versions in a typical *Actinoplanes* sp. SE50/110 cell. However, the number of extrachromosomal copies per cell might be even higher, as it is possible that a proportion of the AICEs was lost during DNA isolation.

3.3.8. Four putative antibiotic production gene clusters were found in the *Actinoplanes* sp. SE50/110 genome sequence

Bioactive compounds synthesized through secondary metabolite gene clusters are a rich source for pharmacologically relevant products like antibiotics, immunosuppressants or antineoplastics [Challis *et al.*, 2000,Hahn & Stachelhaus, 2004]. Besides aminoglycosides, the majority of these metabolites are built up in a modular fashion by using non-ribosomal peptide synthetase (NRPS) and/or polyketide synthase (PKS) as enzyme templates (for a recent review see [Meier & Burkart, 2009]). Briefly, the nascent product is built up by sequential addition of a new element at each module it traverses. The complete sequence of modules may reside on one gene or spread across multiple genes in which the order of the genes is determined by specific linker sequences at the N- and C-terminal ends of their translated proteins [Hahn & Stachelhaus, 2004, Yadav *et al.*, 2003].

For NRPSs, a minimal module typically consists of at least three catalytic domains, namely the andenylation (A) domain for specific amino acid activation, the thiolation (T) domain, also called peptidyl carrier protein (PCP) for covalent binding and transfer and the condensation (C) domain for incorporation into the peptide chain [Hahn & Stachelhaus, 2004]. In addition, domains for epimerization (E), methylation (M) and other modifications may reside within a module. Oftentimes a thioesterase domain (Te) is located at the C-terminal end of the final module, responsible for e.g. cyclization and release of the non-ribosomal peptide from the NRPS [Felnagle *et al.*, 2008].

In case of the PKSs, an acyltransferase (AT) coordinates the loading of a carboxylic acid and promotes its attachment on the acyl carrier protein (ACP) where chain elongation takes place by a β-kethoacyl synthase (KS) mediated condensation reaction [Meier & Burkart, 2009]. Additionally, most PKSs reduce the elongated ketide chain at accessory β-kethoacyl reductase (KR), dehydratase (DH), methyltrans-

ferase (MT) or enoylreductase (ER) domains before a final thioesterase (TE) domain mediates release of the polyketide [Du & Lou, 2010].

In *Actinoplanes* sp. SE50/110, one NRPS (cluster of *Actinoplanes* (cACPL)_1), two PKS (cACPL_2 & cACPL_3) and a hybrid NRPS/PKS cluster (cACPL_4) were found by gene annotation and subsequent detailed analysis using the `antiSMASH` pipeline [Medema *et al.*, 2011]. The first of the identified gene clusters (cACPL_1) contains four NRPS genes (**Fig. 3.17A**), hosting a total of ten adenylation (A), thiolation (T) and condensation (C) domains, potentially making up ten modules. Thereof, seven modules (A-T-C) are entirely located on distinct genes whereas the others are divided by intergenic regions. This suggests an interaction of all four NRPSs in the synthesis of a common product, as only the interaction of all components in the order *nrps1A-B-D-C* leads to the assembly of all domains into ten complete modules with an additional epimerization domain in the last module. These considerations were corroborated by matching linker sequences, named short communication-mediating (COM) domains [Hahn & Stachelhaus, 2004], found at the C-terminal part of NRPS1D and the N-terminal end of NRPS1C. Furthermore, this cluster shows high structural and sequential similarity to the SMC14 gene cluster identified on the pSCL4 megaplasmid from *Streptomyces clavuligerus* ATCC 27064 [Medema *et al.*, 2010]. However, in SMC14 a homolog to *nrps1D* is missing which leads to the speculation that *nrps1D* was subsequently added to the cluster as an additional building block. In fact, leaving *nrps1D* out of the assembly line would theoretically still result in a complete enzyme complex built from nine instead of ten modules. Based on the `antiSMASH` prediction, the amino acid backbone of the final product is likely to be composed of the sequence: Ala-Asn-Thr-Thr-Thr-Asn-Thr-Asn-Val-Ser (**Fig. 3.17A**). Besides the NRPSs, the cluster also contains multiple genes involved in regulation and transportation as well as two MbtH-like proteins, known to be involved in non-ribosomal peptide synthesis [Drake *et al.*, 2007].

The type-1 PKS-cluster cACPL_2 (**Fig. 3.17B**) hosts five genes putatively involved in the synthesis of an unknown polyketide. The sum of the PKS coding regions adds up to a size of ∼49 kb whereas all encoded PKSs exhibit 62-66% similarity to PKSs from various *Streptomyces* strains. However unlike the NRPS-cluster, no cluster structurally similar to cACPL_2 was found in public databases. Analysis of the domain and module architecture revealed a total of ten elongation modules (KS-AT-[DH-ER-KR]-ACP) including nine β-kethoacyl reductase (KR) and eight dehydratase (DH) domains as well as a termination module (TE). However, an initial loading module (AT-ACP) could not be identified in the proximity of the cluster. To elucidate the most likely build order of the polyketide, the N- and C-terminal linker sequences were matched against each other using the software `SBSPKS` [Anand *et al.*, 2010] and `antiSMASH`. Remarkably, both programs independently predicted the same gene order: *pks1E-C-B-A-D*.

Just 15 kb downstream of cACPL_2, a second gene cluster (cACPL_3) containing a long PKS gene with various accessory protein coding sequences could be identified (**Fig. 3.17C**). It shows some structural similarity to a yet uncharacterized PKS gene cluster of S*alinispora tropica* CNB-440 (genes *Strop_2768-Strop_2777*). Besides the

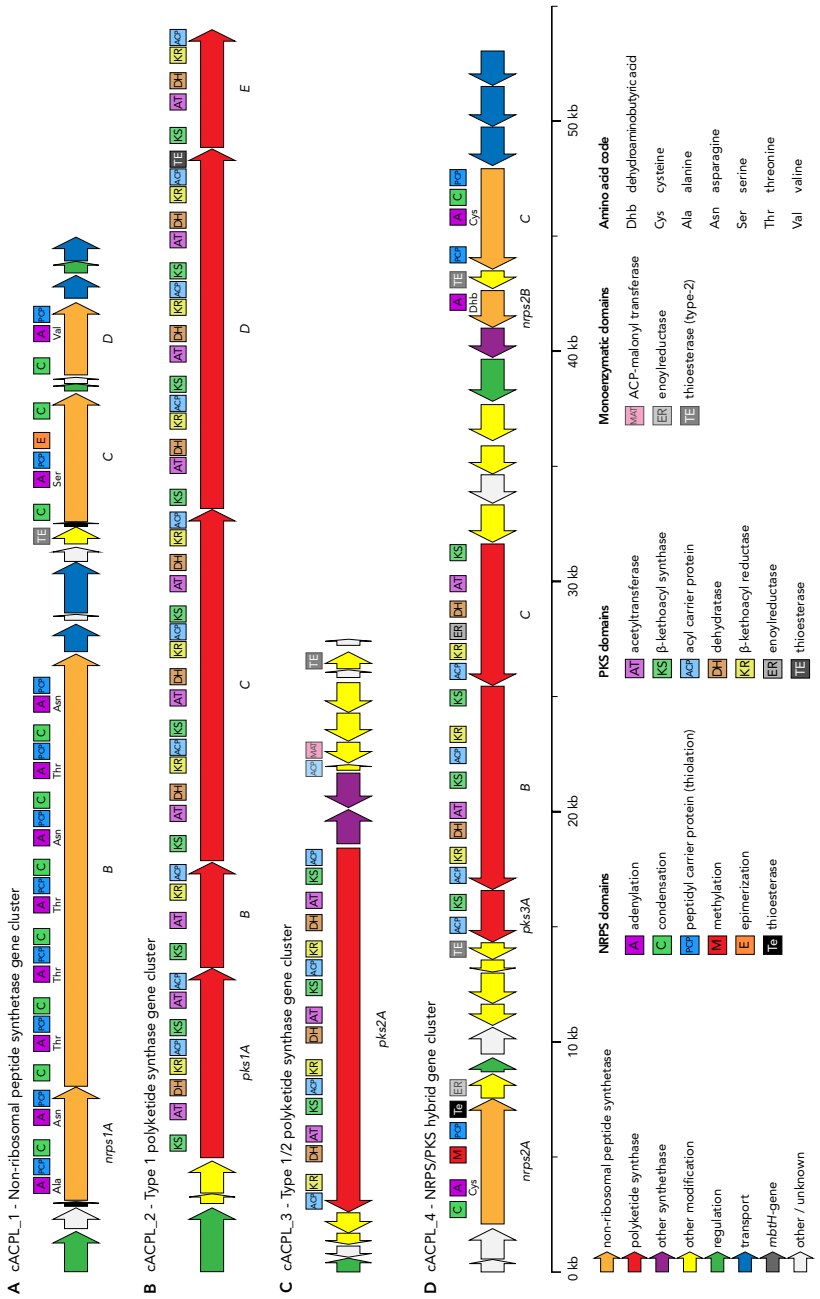

Figure 3.17.: The gene organization of the four putative secondary metabolite gene clusters found in the *Actinoplanes* sp. SE50/110 genome. (**A**) NRPS cluster showing high structural and sequential similarity to the SMC14 gene cluster identified on the pSCL4 megaplasmid from *Streptomyces clavuligerus* ATCC 27064. (**B**) Large PKS gene cluster exhibiting 62-66% similarity to PKSs from various *Streptomyces* strains. (**C**) A single PKS gene with various accessory genes showing some structural similarity to a yet uncharacterized PKS gene cluster of *Salinispora tropica* CNB-440. (**D**) Putative hybrid NRPS/PKS gene cluster with NRPS genes showing high similarity (63-76%) to genes from an uncharacterized cluster of *Streptomyces venezuelae* ATCC 10712 whereas the PKS genes exhibit highest similarity (63-66%) to genes scattered in the *Methylosinus trichosporium* OB3b genome.

three elongation modules identified on *pks2A*, no other modular type-1 PKS genes were found in the proximity of the cluster. However, genes downstream of *pks2A* are likely to be involved in the synthesis and modification of the polyketide, coding for an ACP, an ACP malonyl transferase (MAT), a lysine aminomutase, an aspartate transferase and a type-2 thioestrase. Especially type-2 thioestrases are often found in PKS clusters [Kotowska *et al.*, 2002] like e.g., in the gramicidin S biosynthesis operon [Krätzschmar *et al.*, 1989]. The presence of discrete ACP, MAT and two additional acetyl CoA synthetase-like enzymes is also typical for type-2 PKS systems [Dreier & Khosla, 2000] although no ketoacyl-synthase (KS_α) and chain length factor (KS_β) was found in this cluster [Wawrik *et al.*, 2005].

Another 58 kb downstream of cACPL_3 a fourth secondary metabolite cluster (cACPL_4) was located (**Fig. 3.17D**). It hosts three NRPS and three PKS genes and may therefore synthesize a hybrid product as previously reported for bleomycin from *Streptomyces verticillus* [Shen *et al.*, 2001], pristinamycin IIB from *Streptomyces pristinaespiralis* [Mast *et al.*, 2011] and others [Du *et al.*, 2001]. N- and C-terminal sequence analysis of the two cluster types revealed the gene orders *nrps2B-C-A* and *pks3A-B-C* as most likely. The prediction of the peptide backbone of the NRPS cluster resulted in the putative product dehydroaminobutyric acid (Dhb)-Cys-Cys. One could speculate that the PKSs are used prior to the NRPSs, as *nrps2A* comes with a termination module (Te). However, two additional monomeric thioesterase (TE) domains and one enoylreductase (ER) domain containing genes do also belong to the cluster and may be involved in the termination and modification of the product. Notably, all three NRPS genes show high similarity (63-76%) to genes from an uncharacterized cluster of *Streptomyces venezuelae* ATCC 10712, whereas the PKS genes exhibit highest similarity (63-66%) to genes scattered in the *Methylosinus trichosporium* OB3b genome.

3.4. RNA-sequencing of the *Actinoplanes* sp. SE50/110 transcriptome

In this study, two RNA-seq analysis approaches were carried out. First, a 5'-enriched dataset was used to identify TSS in order to annotate novel protein coding genes, ncRNAs, and antisense transcripts. Based on this information, gene start site corrections and other annotation improvements were performed. Second, a full-length transcript dataset (non-5'-enriched) was used to measure transcript expression values and to perform differential expression testing between different *Actinoplanes* sp. SE50/110 cultivations. The individual steps of this analysis strategy are depicted in **Figure 1.6** and described in detail within **Sections 3.4.2** and **3.4.3**.

The transcriptome analysis for *Actinoplanes* sp. SE50/110 was carried out using RNA-sequencing technology because of the availability of the full reference genome sequence and its methodological advantages over standard microarrays. While microarrays need to be specifically designed for the organism under investigation and are afflicted with several disadvantages such as saturation effects and background noise, RNA-seq experiments can be conducted without a specific design and yield high quality sequence data on a base-pair resolution [Wang *et al.*, 2009]. These

benefits combined with novel analysis methods enhance the popularity of the technology, which is best reflected by its diverse application in bacteria [Güell *et al.*, 2009, Sharma *et al.*, 2010], archaea [Wurtzel *et al.*, 2010], yeast [Nagalakshmi *et al.*, 2008, Yuan *et al.*, 2011], plants [Lister *et al.*, 2008, Massa *et al.*, 2011], and mammals [Mortazavi *et al.*, 2008, Esteve-Codina *et al.*, 2011] including man [Cloonan *et al.*, 2008, Chen *et al.*, 2011a].

Up to now, practically nothing is known about gene expression and regulation in *Actinoplanes* sp. SE50/110. However, it is generally known that growth in different cultivation media leads to changes in gene expression, influencing the productivity of a strain to a great extend [Lee *et al.*, 1997, Jung *et al.*, 2008]. Correspondingly, a variety of cultivation media were used for growing *Actinoplanes* sp. SE50/110 in the past, resulting in acarbose yields between 0 and 1 g/L [Rauschenbusch & Schmidt, 1978]. By conducting further cultivation experiments, it was shown that maltose containing media induce acarbose production, whereas glucose has a negative effect on its production rate [Brunkhorst & Schneider, 2005, Wang *et al.*, 2011a]. Despite these insights, the underlying changes in gene expression remain concealed. In order to uncover these changes, three different growth media for cultivation and transcriptome analysis of *Actinoplanes* sp. SE50/110 were selected in this work. First, a defined minimal medium (Mal-MM) with maltose as sole carbon source was used to serve as a reference for a reliable acarbose production level (**Tab. 2.3**). Second, the same medium with supplemented trace elements (Mal-MM-TE) was used to study the impact of trace elements on growth rate and acarbose production efficacy (**Tab. 2.4**). Third, a complex medium (Glc-CM) with glucose as main carbon source was utilized to serve as a non-producing counterpart in order to study the expressional changes between acarbose inducing and acarbose repressing media (**Tab. 2.2**).

3.4.1. Cultivation of *Actinoplanes* sp. SE50/110 for transcriptome analysis

Actinoplanes sp. SE50/110 was grown in the three different cultivation media. In order to compare cultivation results, cell dry weights (CDWs) were determined by weighing the pellets after centrifugation and freeze-drying; acarbose concentrations were determined by HPLC and UV-detection (see Materials and Methods **Sections 2.6.4** and **2.6.5**). **Figure 3.18** shows the CDW and the acarbose production of the three conditions at day four of the cultivation (early stationary phase). As expected, no acarbose could be detected in the supernatant of the Glc-CM condition, whereas moderate levels of acarbose were detected in both Mal-MM media. Interestingly, the supplied trace elements increased the acarbose production by 50% and the growth in terms of CDW by 42%. The similar increase in both observed parameters suggests a linear correlation between the number of cells and the amount of acarbose produced, leading to the hypothesis that the supplied trace elements mainly promote cell growth, which in turn causes an indirect increase in acarbose yields (opposed to a production increase on a per cell basis). This could be confirmed when the acabose yields were normalized to the cell dry weight of the cultures, resulting in 44 mg per gram CDW for Mal-MM and 46 mg per gram CDW for Mal-MM-TE.

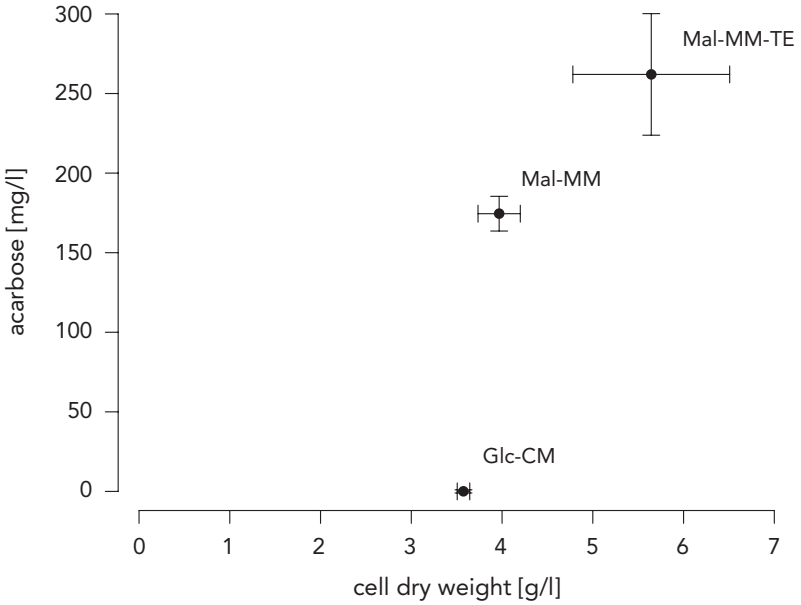

Figure 3.18.: Cell dry weight and acarbose production of *Actinoplanes* sp. SE50/110 cultures grown in three different media. *Actinoplanes* sp. SE50/110 was grown in maltose containing minimal medium (Mal-MM), Mal-MM with trace elements (Mal-MM-TE) and glucose containing complex medium (Glc-CM). After four days of cultivation, samples were taken and analyzed for cell dry weight (CDW) and acarbose production yields. Mean values of biological replicates are shown, the bars indicate the standard deviation in both variables.

The observed induction of acarbose production by maltose and its repression by glucose containing media are in good accordance with the literature [Brunkhorst & Schneider, 2005, Wang *et al.*, 2011a]. The conditions were therefore well suited for subsequent RNA-seq analysis aimed at the identification of differentially expressed (DE) genes.

3.4.2. Improving the *Actinoplanes* genome annotation by RNA-seq

In order to improve the genome annotation of *Actinoplanes* sp. SE50/110, three 5'-enriched cDNA libraries were constructed and sequenced. Each library was based on isolated RNA that was pooled after extraction from all biological replicates from the three cultivation conditions (4× Mal-MM, 4× Mal-MM-TE, and 2× Glc-CM). The RNA isolation was carried out using the TRIzol (Life Technologies) and the RNeasy Mini Kit (QIAGEN) as described in the Materials and Methods **Section 2.6.3**. Terminator exonuclease (TEN) treatment was used to digest stable RNA and yield 5'-enriched fragments. The cDNA library preparation was carried out with the help of the TruSeq RNA sample prep Kit (Illumina). Each of the three prepared libraries were then loaded on one lane and sequenced on an Illumina GA IIx platform. About 9 million reads were sequenced in total of which ∼1 million passed subsequent

strict quality filtering and mapped to the reference genome (**Tab. 3.8**). Thereof, 703,462 reads mapped unambiguously and were used for the identification of TSSs. It was also found that, on average, a mapped read (readlength 36 bp) matched 2.35 times on the genome sequence (see **Table 3.8**, column maprate), which was mainly caused by reads aligning to the six *rrn* operons of *Actinoplanes* sp. SE50/110.

Table 3.8.: RNA-sequencing results of 5'-enriched libraries used for annotation improvement.

Condition	Sequenced reads	Mappable reads	Unique matches	Maprate
Mal-MM	7,889,721	810,404 (10.27%)	553,311 (7.01%)	2.25×
Mal-MM-TE	932,245	158,447 (17.00%)	115,659 (12.41%)	2.19×
Glc-CM	220,625	109,589 (49.67%)	34,492 (15.63%)	3.34×
Total	9,042,591	1,078,440 (11.93%)	703,462 (7.78%)	2.35×

Based on the mapped data, the applied strategy for genome annotation improvement by RNA-seq is summarized in the following seven steps, which are then elaborated in detail in the subsequent sections.

1. Detect all TSSs and determine their local genomic context.
2. Correct the translational start codon of protein coding genes where a TSS clearly indicates wrong automatic annotation.
3. Derive consensus -10 and -35 motifs for the promotor regions of *Actinoplanes* sp. SE50/110 genes.
4. Search and annotate longer unoccupied TSS downstream regions for putative novel CDSs.
5. Search and annotate also shorter unoccupied TSS downstream regions for putative novel ncRNAs.
6. Inspect the remaining TSS regions that exhibit a -10 (and optionally -35) region and annotate them as putative ncRNAs with unknown function.
7. Annotate antisense transcripts for all genes.

Detection of Transcription start sites

The mapping results were analyzed for aggregated stacks of 5'-enriched reads that constitute putative TSSs. Because transcription usually starts at a distinct base, the consequential sudden increase of coverage is used to infer the exact position of the TSS [Knippers, 2001]. In more detail, the difference between the coverages of the last base before, and the first base of the TSS is considered to be the Δ *stacksize* of the TSS as exemplified in **Figure 3.19**. By applying a threshold for the Δ stacksize, it is possible to control the sensibility and specificity of the method. After empirical tests with various thresholds, ten was chosen as the cutoff for this dataset, as it showed the highest specificity after manual inspection of randomly selected TSSs from the result set.

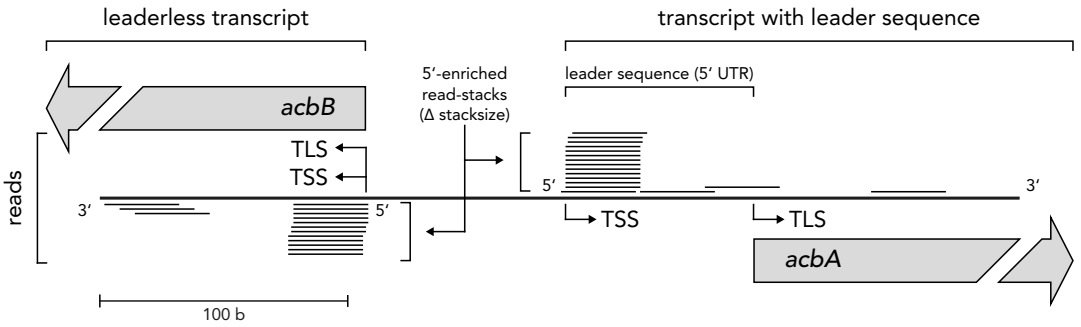

Figure 3.19.: Scheme of the detection mechanism for transcription start sites (TSSs) using RNA-sequencing of 5'-enriched cDNA libraries. The excerpt shows the detailed positions of 5'-enriched reads that map to the genes *acbB* and *acbA* from the acarbose gene cluster. The differences in reads per base are scanned throughout the genome and putative TSS positions are reported if two adjacent bases exhibit a Δ in the stacksize that is above a given threshold. The gene *acbB* possesses a leaderless TSS whereas *acbA* owns a clear leader sequence between the TSS and the translation start (TLS) of the coding sequence. The depicted data was taken from the Mal-MM condition; reads are 36 bases in length.

In total, 1427 putative TSSs were detected by this procedure. Subsequent filtering of the results yielded 799 TSSs that did not overlap with upstream CDSs or precede RNA genes. The filtered set was then analyzed for potential correlations between TSS coverage and distance to the next downstream translation start (TLS) of a gene (**Fig. 3.20**). The analysis showed an accumulation of TSSs between 10-500 bases upstream of TLSs as well as a stacksize between 10-100 reads per base. Manual inspection of TSSs that were located more than 500 bases away from the TLSs revealed an increasing amount of putatively unannotated genes whose CDSs were mostly shadowed by questionable overlapping annotations on the complementary strand. However, in many cases the annotations seemed to be correct and the TSS might initiate antisense transcription of these genes as described later.

Based on these observations, the distances of 661 filtered TSSs that resided within 500 bp upstream of annotated genes were analyzed and revealed a ratio of ∼20% leaderless transcripts to ∼80% transcripts that provide a 5'-untranslated region (UTR) in *Actinoplanes* sp. SE50/110. The length of the 5'-UTR varies in size but shows a peak around 35 bp length (**Fig. 3.21**).

Gene start correction using RNA-seq

About a third of all TSSs did overlap with CDSs and were analyzed in this section. These TSSs can be used for the correction of premature CDS starts in cases where they are located shortly after the original start codon. For the reason that a start codon must occur after the TSS, the correction process involves the new annotation of the next in-frame start codon after the TSS. **Figure 3.22** gives an overview of the positions of the 438 TSSs within the corresponding CDSs. From this histogram it is ev-

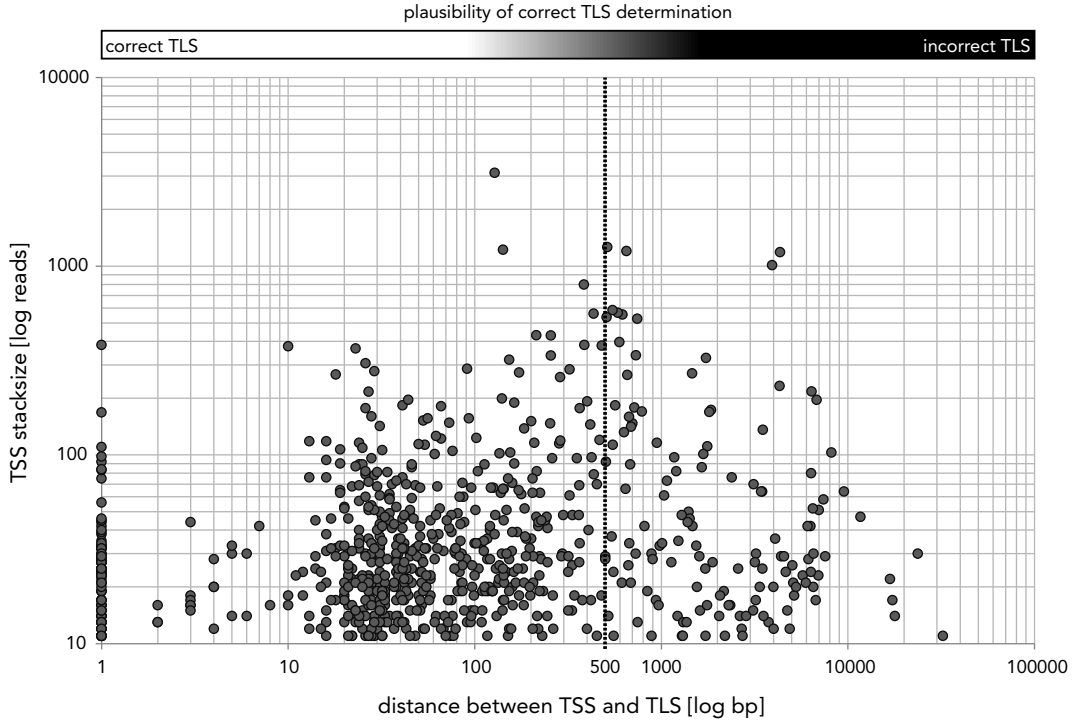

Figure 3.20.: Scatterplot of transcription start site (TSS) coverage and distance to next downstream translation start (TLS). The figure shows the coverage of 799 TSSs plotted against the distances to the next downstream TLSs. TSSs were excluded when they were followed by RNA genes or when they did overlap upstream CDSs. A threshold of 500 bp distance to the TLS is marked by the vertical dashed line.

ident that about 30% of the TSSs fulfill these requirements, i.e. they are located within the first 10% of the CDSs, and pose candidates for translation start site correction. After manual inspection of the 126 candidates, 41 CDS starts were unambiguously found to be wrongly annotated and were subsequently corrected (**Tab. A.1**). Most of the other candidate TSSs were accompanied by at least one other TSS that was correctly located upstream of the corresponding CDS. As the additional identified stacks might represent valid alternative TSSs for these genes, the original longer CDS annotation was not changed.

Although the TSS identification parameter (Δ stacksize of ten) and the TLS correction threshold ($< 10\%$) were chosen rather conservatively, remarkably few CDSs had a clearly erroneous annotation.

Promotor element identification using RNA-Seq

Based on the knowledge of the exact TSS positions, upstream and downstream regions were analyzed next for possible conserved promotor elements, such as the Pribnow box (-10 region) and the -35 region. For this analysis, the 135 leaderless transcripts iden-

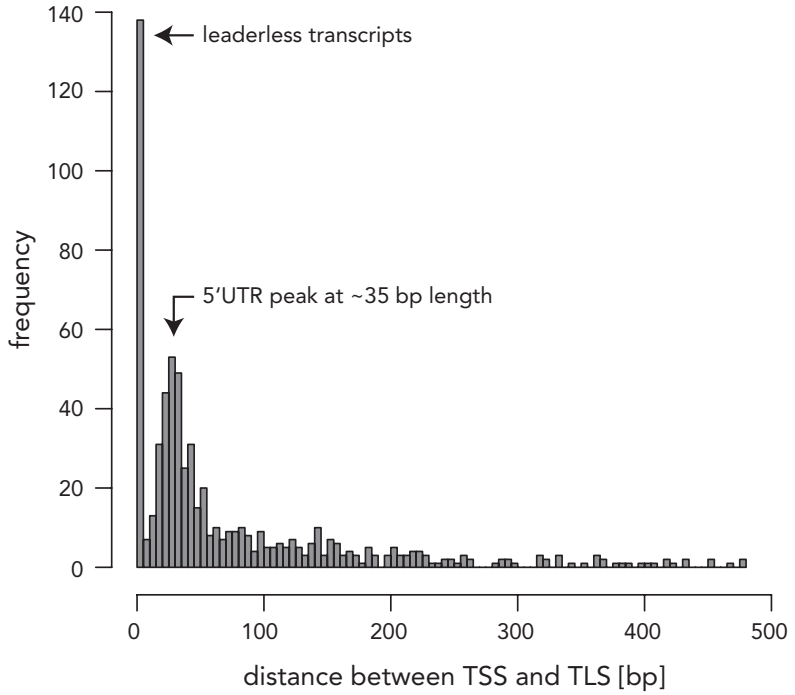

Figure 3.21.: Histogram of distances between 661 transcription start sites (TSSs) and the next downstream translation starts (TLSs). The initial peak holds 135 leaderless transcripts whereas the remaining bars sum up to 526 transcripts with varying 5'-untranslated regions (UTRs) length. The histogram bucket size is 5 bp.

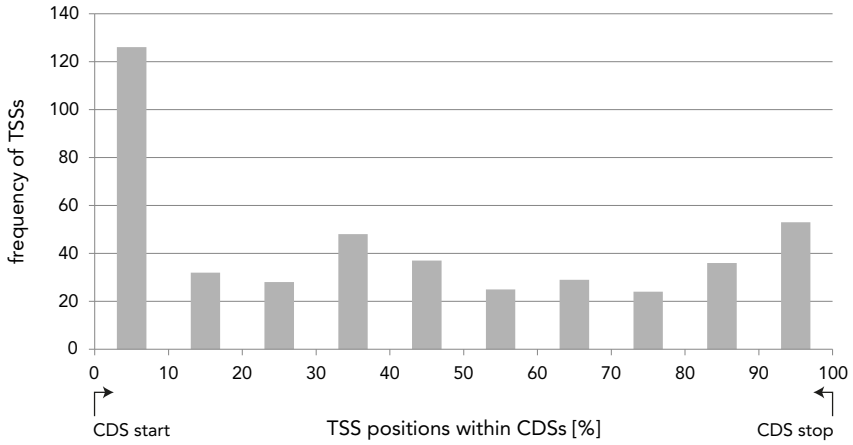

Figure 3.22.: The histogram shows the positions of 438 transcription start sites (TSSs) that did overlap coding sequences (CDSs) on the sense strand. Each bar corresponds to 10% length of the underlying CDSs.

tified in **Section 3.4.2** were used first, as their alignment was expected to be most conserved. **Figure 3.23A** shows the resulting 50 bases consensus region (40 bases upstream and 10 bases downstream) of this analysis involving 135 leaderless transcripts calculated by means of the `WebLogo` software [Crooks *et al.*, 2004]. As expected, the start codon is highly conserved at positions 1-3. Interestingly, the most conserved upstream base is a cytosine at position -1. Furthermore, a degenerated -10 region can be assumed from positions -12 to -7 with the consensus sequence `(A/T)ANNNT`. Overall, the prevalence of guanine and cytosine bases – caused by the high GC-content – can also be clearly observed. However, no signs of a -35 region could be identified with this method.

A similar analysis was performed using 413 transcripts with a 5'-UTR sequence in order to identify possible differences in the promotor regions between leaderless and leader transcripts of *Actinoplanes* sp. SE50/110. The 413 transcripts exhibit a 5'-UTR length ranging from 3-100 bases, which implies the exclusion of leaderless transcripts as well as translation start codons at positions 1-3 of the consensus. As evident from **Figure 3.23B**, the consensus sequence of this analysis shows a less conserved but identical -10 region. Interestingly, the -1 cytosine is still present in an equally conserved manner in conjunction with a `G/A` and `T/A` at positions +1 and +2, respectively. Also, the -5 guanine present in leaderless upstream regions is not found to be conserved in transcripts with a leader sequence.

In order to refine the Pribnow box consensus pattern for *Actinoplanes* sp. SE50/110, a more sophisticated analysis method was applied next that allowed a variable -10 region positioning for better detection. The tool `PRISM` [Carlson *et al.*, 2006] identified the -10 consensus motif `TANNNT` in 62.4% of the leaderless transcripts and in 56.9% of the transcripts with leader sequences. The motif resembles the consensus sequence of the *E. coli* σ-70 protein recognition site `TATAAT` in the three higest conserved bases [Singh *et al.*, 2011]. Its 5'-end was located at position -12.4 on average (**Fig. 3.24**). Additionally, a putative -35 consensus motif was identified in 19.3% of the examined TSS upstream regions starting at a mean position of -35.0. Its sequence `(G/A/T)NTT(G/T)(C/A)` seems to partially overlap with the -35 consensus motif `TTGACA` of *E. coli* but is obviously less conserved (consensus overlap `TTga`). The distance between both promotor elements was found to be 17.6 bases on average which is very close to 17, the optimal spacing found for these elements in *E. coli* [Singh *et al.*, 2011]. The consensus promotor recognition elements for a σ-70 protein homologue in *Actinoplanes* sp. SE50/110 are visualized in **Figure 3.24**.

Identification of novel CDSs by RNA-seq

According to the 5'-UTR length distribution of normal transcripts (**Fig. 3.20**), 5'-UTRs with more than 500 bp length are unlikely to belong to correctly annotated genes. Therefore, the downstream regions of these TSSs are promising targets for finding novel CDSs which were not reported by the automatic annotation pipeline `GenDB`. For this analysis it was important to not only consider the sense strand but also the antisense strand, as CDS regions usually do not overlap each other. Based

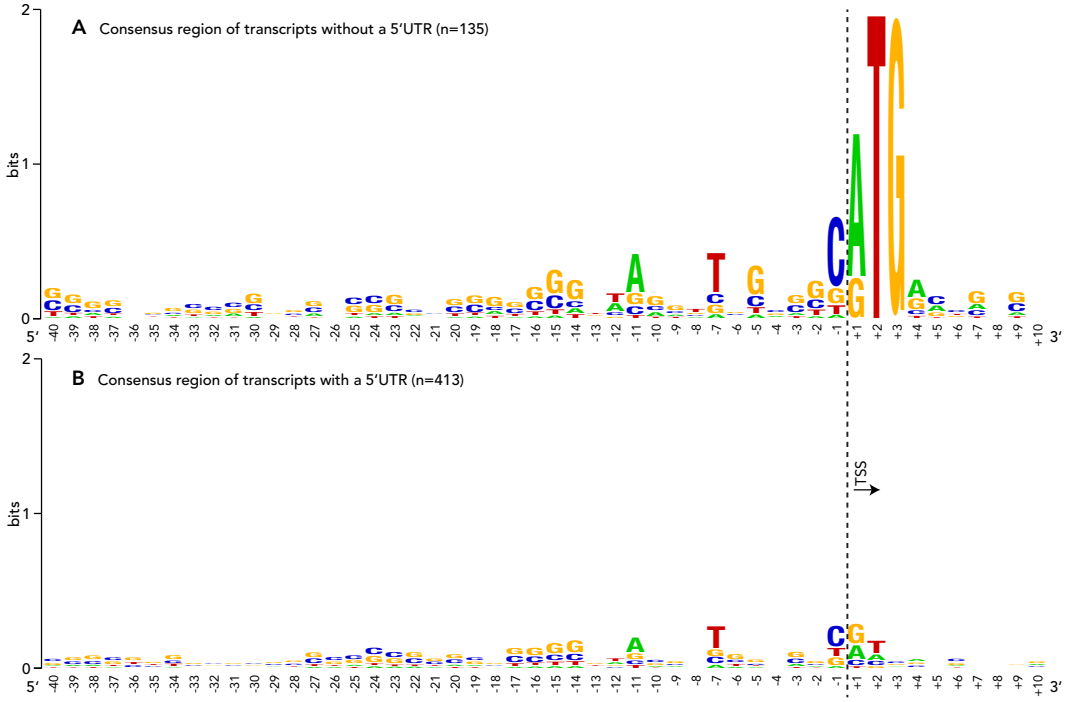

Figure 3.23.: Promotor consensus regions of *Actinoplanes* sp. SE50/110 transcription start sites (TSSs). The figure shows a 50 b window around the TSSs of (**A**) 135 leaderless transcripts (transcripts with no 5'-UTR) and (**B**) 413 transcripts that exhibit a 5'-UTR between 3-100 b in size. The 50 b window is subdivided into a 40 b upstream region, and a 10 b downstream consensus sequence. The images were created using the WebLogo software [Crooks *et al.*, 2004]. The higher a specific nucleotide base is conserved within the region, the larger is its representation in the illustration.

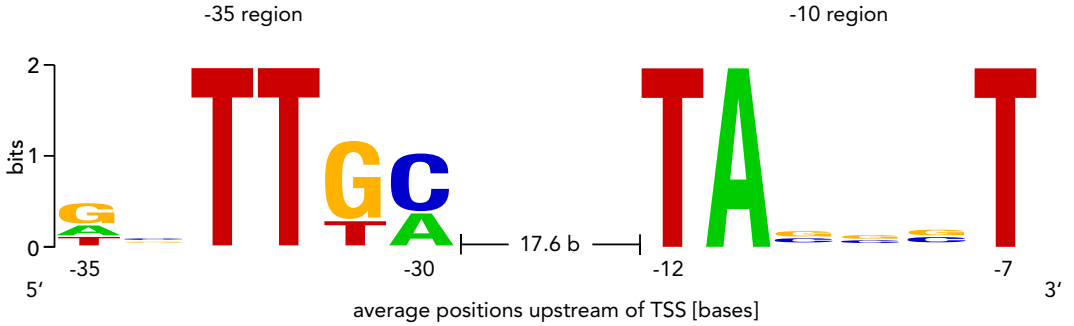

Figure 3.24.: The image shows the -10 and -35 consensus recognition motifs identified for *Actinoplanes* sp. SE50/110 σ-70-like proteins in *Actinoplanes* sp. SE50/110. The depicted positions are averaged from 103 occurrences of the -35 region and 286 occurrences of the -10 region. Values are relative to the TSS. The motifs were identified using the PRISM software [Carlson *et al.*, 2006].

on this restriction, only 41 TSSs were found to be followed by an unoccupied 500 bp downstream region on both strands. In order to increase the number of searchable sequences the 500 bp limit was relaxed to 100 bp, which yielded 249 target regions between 101 and 1388 bp in length.

For the reason that no genes were annotated in these sequences it was assumed that putative unannotated CDSs may be exceptional in terms of length and/or base composition, which was derived from the observation that the average GC-content of the 249 sequences was only 66.0%. To account for these atypical CDSs, a second gene prediction software besides `Prodigal` was applied. The gene finder `GeneMarkS` was chosen for this task because of its iterative self-training algorithm that does not rely on previous knowledge about the sequences. More importantly, it utilizes a positional nucleotide frequency model that may better cope with the anticipated atypical sequence composition. It was also shown that `GeneMarkS` has a robust performance in identifiying small and atypical CDSs, which seems appropriate for this analysis [Besemer *et al.*, 2001].

In point of fact, `GeneMarkS` predicted eight novel CDSs whereas `Prodigal` reported only three putative CDSs (**Tab. 3.9**). Interestingly, just one of the CDSs, *acpl8397*, was predicted by both programs and, at the same time, showed high sequence similarity to integrases from *Streptomyces zinciresistens* K42 and *Streptomyces coelicolor* A3(2). The other predictions had very poor or no sequence similarity at all to protein sequences from public databases with the exception of *acpl8401*, which showed a good similarity to an unnamed protein from *Salinispora arenicola* CNS-205.

Table 3.9.: Novel CDS predicted in TSS downstream regions based on RNA-seq data.

Gene	Strand	CDS start	CDS stop	CDS length	Gene finder	Description
acpl8395	+	108298	108477	180	GeneMarkS	hypothetical protein
acpl8402	+	180576	180704	129	GeneMarkS	recombinase domain
acpl8401	+	916415	916606	192	GeneMarkS	unnamed protein
acpl8400	-	6319054	6318917	138	GeneMarkS	hypothetical protein
acpl8399	-	6631580	6631296	285	GeneMarkS	hypothetical protein
acpl8398	-	6869314	6869177	138	GeneMarkS	hypothetical protein
acpl8397	-	7188476	7187919	558	both	integrase
acpl8404	-	7655920	7655504	417	Prodigal	hypothetical protein
acpl8403	-	7955707	7955522	186	Prodigal	hypothetical protein
acpl8396	+	7974216	7974629	414	GeneMarkS	hypothetical protein

Moreover, the amino acid translation of the short novel gene *acpl8402* showed almost perfect identity to subsequences of two other genes from *Actinoplanes* sp. SE50/110, namely the tyrosine recombinase gene *acpl263* and the integrase family protein coding gene *acpl299*. A multiple sequence alignment together with the two other tyrosine recombinases Acpl340 and Acpl283 revealed that the C-terminal ends of these proteins were well conserved (**Tab. 3.10**). The new Acpl8402 protein sequence

aligned nicely to this region although no conserved catalytic domain was located in this area according to the conserved domain database (CDD) [Marchler-Bauer *et al.*, 2011b]. Even more intriguing, this region had no similarity to other public protein sequences which together leads to the speculation that it fulfills a specific task in *Actinoplanes* sp. SE50/110 and might function independently as in Acpl8402 or fused to intergrase/recombinases. For the reason that these enzymes act on DNA sequences, an obvious function might be the binding of DNA, which could also be independently used e.g. as in DNA-binding regulatory proteins.

Table 3.10.: Multiple sequence alignment of *Actinoplanes* sp. SE50/110 intergrase/recombinase C-terminal ends.

```
Acpl263    365  SSAVTTADTYWTVFRELADRAVAATAGLLR-------THARIRLNLGAASQA-  436
Acpl8402     0  ---VTTADTYWTVFRELADRAVTATAGLLR-------SHARIRLNLGAASQA-   42
Acpl299     25  SSAVTTADTYWTVFRELAHQAVAVTAGLLR-------THARFRLRLEAASQA-   96
Acpl340    363  TSYAFTADTYATVLPDQAKHAAESTARLVLDALNEARPAVGARLGPGLATASS   442
Acpl283    475  TSYAFTADTYATVLPDQAKHAAESTARLVLNALHKACTAAGA----GSQTGS-  549
                   . ***** **: : *.:*.  ** *:        . .         : :
```

Identification of non-coding RNAs

Downstream regions of TSSs are promising targets for identifying ncRNAs by searching these sequences against RNA databases, such as `Rfam` [Griffiths-Jones, 2004, Gardner *et al.*, 2009], `fRNAdb` [Kin *et al.*, 2007, Mituyama *et al.*, 2009], and `NONCODE` [Liu, 2004, He *et al.*, 2008]. Performing these searches resulted in the clear identification of nine ncRNAs with known functions (**Tab. 3.11**) of which four are briefly described in the following paragraphs.

Table 3.11.: Identified non-coding RNAs with known function.

Gene	Gene Symbol	Gene length	Description
acpl8386	ssrA	384	transfer-messanger RNA
acpl8388	rnpB	404	ribonuclease P class A RNA
acpl8389	cobRS	179	cobalamin riboswitch RNA
acpl8392	selC	92	selenocysteine transfer RNA
acpl8390		98	signal recognition particle RNA
acpl8391		119	SAM riboswitch (S box leader) RNA
acpl8393		111	thiamine pyrophosphate (TPP) riboswitch RNA
acpl8394		57	msiK RNA
acpl8387		72	6C RNA

The transfer-messanger RNA The first of the identified ncRNAs was a transfer-messenger RNA (tmRNA), which is one component of the ribonucleoprotein complex that is responsible for resetting ribosomes that were stalled during translation because

of erroneously transcribed mRNAs [Keiler, 2008]. The newly annotated gene *ssrA* overlaps an adjacent recombinase encoding gene *acpl1084* by 156 bp (**Fig. 3.25**). Furthermore, the ribonucleoprotein complex consists of three other components, a 'small protein B' (SmpB), an 'elongation factor thermo unstable' (EF-Tu), and a ribosomal protein S1 (RPS1). Interestingly, the *smpB* homologue was also found in close proximity to *ssrA*, only separated by *acpl1083*, which putatively encodes a RNA polymerase subunit with partial sequence similarity to a RNA polymerase σ-factor from *Frankia* sp. CN3. The other depicted genes encode a pyruvate decarboxylase isozyme 2 (PDC2) and a DNA translocase (FtsK). In *E. coli*, tmRNA is one of the most abundant types of RNA in the cell which correlates nicely with the extreme expression observed for the *Actinoplanes* homologue (**Fig. 3.25**). Moreover, significant amounts of antisense transcripts were found for the *ssrA* gene, although these are still 1-2 orders of magnitude less abundant than the main transcripts.

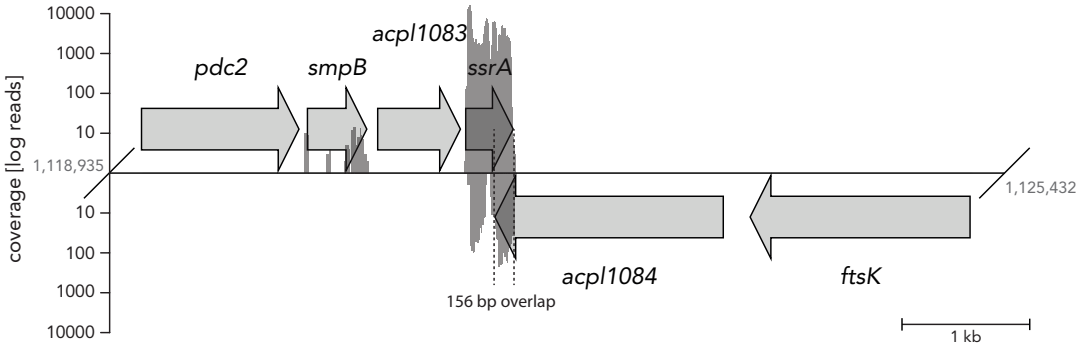

Figure 3.25.: Genomic vicinity of the transfer-messenger RNA (tmRNA) gene *ssrA* of *Actinoplanes* sp. SE50/110. The gene overlaps the adjacent recombinase encoding gene *acpl1084* by 156 bp. The other depicted genes encode a pyruvate decarboxylase isozyme 2 (PDC2), a 'small protein B' (SmpB), a putatively RNA polymerase σ-factor (Acpl1083), and a DNA translocase (FtsK). The y-axis shows the coverage of 5'-enriched reads in this region of the genome.

The ribonuclease P RNA A second identified ncRNA constitutes a ribonuclease P RNA (RNase P), which is responsible for processing various RNAs, including its preferred substrate, precursor-tRNA, where it cleaves the 5'-leader element off all nascent tRNAs [Hartmann *et al.*, 2009]. In bacteria, the ribozyme is accompanied by a single essential protein (termed *C5*), which increases the substrate range, reaction rate, and assists in the release of the product from the holoenzyme [Sun *et al.*, 2006]. While the gene of protein C5, *rnpA*, was already identified through conventional genome annotation (*acpl8384*) only 1.5 kb away from the *oriC*, the novel RNA gene *rnpB* was located 1.6 Mb apart from its protein subunit (**Fig. 3.26**). Furthermore, the lengths of both subunits (404 bases and 119 aminoacids) agree well with the sizes of corresponding genes found in other bacteria [Sun *et al.*, 2006]. Interestingly, the protein

C5 did not exhibit an own distinct TSS, which suggests its co-transcription with the adjacent gene *rpmH*, encoding a 50S ribosomal subunit protein L34 (**Fig. 3.26**).

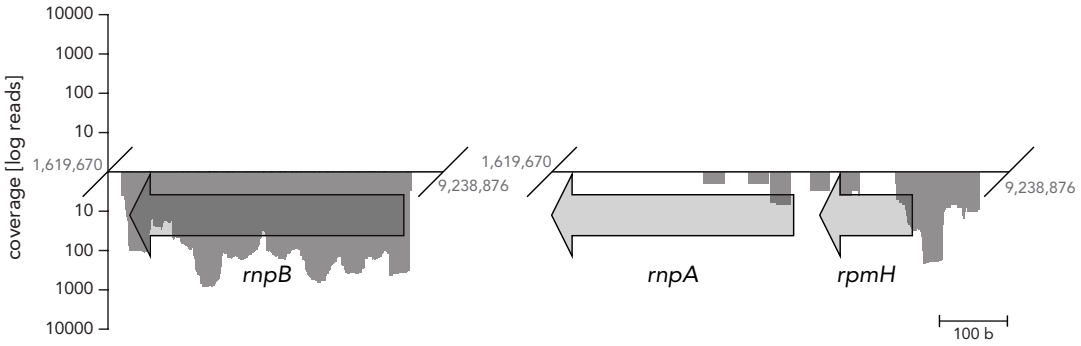

Figure 3.26.: Genomic vicinities of the RNase P gene *rnpB* and its associated protein C5 coding gene *rnpA*. The adjacent gene *rpmH* encodes a 50S ribosomal subunit protein L34 and forms a putative bicistronic operon with *rnpA*.

The cobalamin riboswitch RNA The third ncRNA exhibits high similarity to bacterial cobalamin riboswitch RNAs, which act as cis-regulatory elements in the 5'-UTRs of cobalamin (vitamin B12) related genes [Nahvi *et al.*, 2002]. In more detail, the riboswitch changes its conformation in the presence of its effector adenosylcobalamin (Ado-CBL) which leads to the folding of an adjacent regulatory structure that represses the transcription of vitamin B12 related genes [VITRESCHAK *et al.*, 2003]. Interestingly, this cobalamin riboswitch was identified in the 5'-UTR of the bicistronic operon *nrdLM* encoding the two subunits of a ribonucleotide reductase (RNR), an essential enzyme that provides the building blocks for DNA synthesis and repair in all living cells [Reichard, 1993]. Two types of RNRs were identified in *Actinoplanes* sp. SE50/110. Class I RNR contains two subunits R1 (α_2) and R2 (β_2) encoded by *nrdL* and *nrdM*, which form an oxygen dependent and cobalamin independent tetrameric enzyme complex of two R1 and two R2 subunits (**Fig. 3.27**). In contrast, the class II RNR consist of an oxygen independent and cobalamin dependent homodimer encoded by *nrdE* [Torrents *et al.*, 2002].

The cobalamin riboswitch in the 5'-UTR of the class I RNR is therefore likely to repress the transcription of RNR in the presence of vitamin B12 as was shown for a homologous system in *S. coelicolor*, where the class II RNR is the primary system for deoxyribonucleotide synthesis [Borovok *et al.*, 2006]. Similar to *S. coelicolor*, the class II RNR operon of *Actinoplanes* sp. SE50/110 contains a second gene, encoding the putative transcriptional repressor NrdR (**Fig. 3.27**). NrdR was shown to repress both RNR systems in *S. coelicolor* by binding to a repeat motif upstream of their promotor regions [Borovok *et al.*, 2004]. On the other hand, *Actinoplanes* sp. SE50/110 lacks an AraC-like regulatory protein encoding gene *nrdS* that is present in the class I RNR operon of *S. coelicolor*.

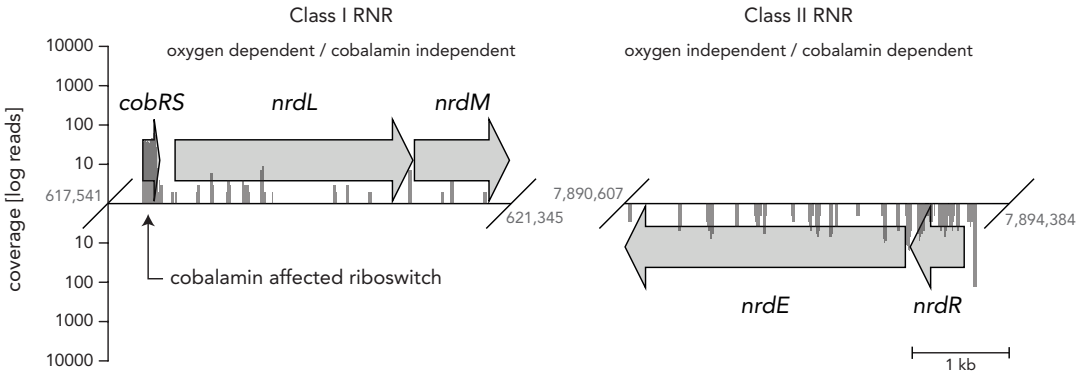

Figure 3.27.: Genomic vicinities of both ribonucleotide reductase (RNR) clusters of *Actinoplanes* sp. SE50/110. The class I operon encodes the two subunits of the heterotetrameric enzyme and is controlled by the newly discovered cobalamin (vitamin B12) affected riboswitch RNA. The second operon encodes the class II homodimeric RNR NrdE and a putative transcriptional regulator NrdR.

The selenocysteine transfer RNA A fourth newly discovered ncRNA represents a selenocysteine-specific transfer RNA ($tRNA^{Sec}$), which was not discovered by the `tRNAscan-SE` software. Selenocysteine is the 21. proteinogenic amino acid and is essential to a variety of important proteins. Most of these *selenoproteins* have redox activities, such as the formate dehydrogenase, the glutathione peroxidase, and the glycine reductase [Kryukov & Gladyshev, 2004].

The pathway by which selenocysteine is synthesized and incorporated into nascent selenoproteins is encoded by four genes, *selA* (selenocysteine synthase), *selB* (selenocysteine-specific elongation factor), selC (selenocysteine-specific tRNA), and selD (selenophosphate synthetase). It starts by acetylation of $tRNA^{Sec}$ (SelC) with serine, which is then bound by selenocysteine synthase (SelA) together with the selenium donor molecule selenophosphate. Selenophosphate is synthesized by the selenophosphate synthetase (SelD) from selenide and adenosine triphosphate (ATP). SelA then catalyzes the conversion of selenophosphate and serinyl-$tRNA^{Sec}$ to selenocysteinyl-$tRNA^{Sec}$, which is subsequently released from SelA and bound to the elongation factor SelB. This complex requires the presence of an `UGA` codon in the selenoprotein mRNA in conjunction with a special mRNA secondary structure element, which together, finally leads to the incorporation of selenocysteine into the nascent polypeptide chain [Gursinsky *et al.*, 2000].

With the discovery of the *selC* gene, the complete gene set is identified in *Actinoplanes* sp. SE50/110 (**Fig. 3.28**). In contrast to other organisms, the selenocysteine biosynthesis cluster is divided by three presumably unrelated genes, encoding an acyltransferase 3 (*acpl5749*) and two hypothetical proteins (*acpl5747* and *acpl5748*) [Gursinsky *et al.*, 2000].

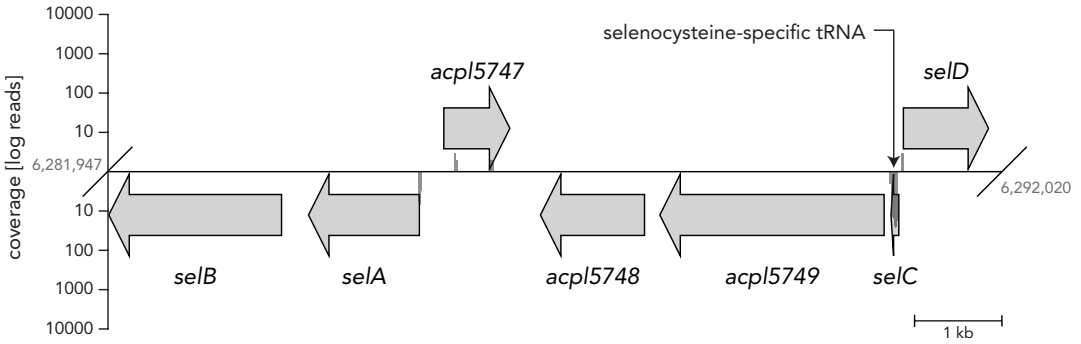

Figure 3.28.: Genomic vicinity of the selenocysteine biosynthesis gene cluster of *Actinoplanes* sp. SE50/110. The involved genes encode selenocysteine synthase (*selA*), a selenocysteine-specific elongation factor (*selB*), the newly discovered selenocysteine-specific tRNA (*selC*), and a selenophosphate synthetase (*selD*). The cluster is separated by three putatively uninvolved genes, encoding an acyltransferase 3 (*acpl5749*) and two hypothetical proteins (*acpl5747* and *acpl5748*).

Annotation of novel non-coding RNAs with unknown function

The remaining ncRNAs for which no function could be determined by databases searches were further analyzed for their probability to constitute novel ncRNAs with yet unknown function. Only TSSs were analyzed that exhibited at least an upstream -10 motif in the promotor region as derived from **Section 3.4.2** and at least 250 bases of unoccupied downstream sequence. The latter restriction was necessary to prevent the annotation of TSSs that more likely belong to already known downstream genes. In total, 39 potential ncRNAs were determined by this procedure.

However, the detection of previously unknown ncRNAs is a challenging task because ncRNAs lack the usage of codons, which eases the detection of CDSs for protein coding genes to a great extend. Moreover, ncRNAs are generally less conserved than mRNAs because mutations can not lead to frameshifts and, thus, are only relevant when they occur in active sites of the RNA gene [Pang *et al.*, 2006]. While a TSS can overtake the function of a start codon in determining the ncRNA start, no stop codon equivalent is available. Hence, the length of a ncRNA depends on the size of its transcript, which can not be determined with a 5'-enriched cDNA library alone. Nevertheless, as described in **Section 3.4.3**, another RNA-seq run with a suitable library was also performed in this work, which was utilized to determine the approximate lengths of the novel ncRNAs.

After manual inspection of the 39 candidates, 18 most likely ncRNAs were annotated in the *Actinoplanes* sp. SE50/110 genome (**Tab. A.3**).

Annotation of antisense RNAs

After the annotation of all potential ncRNAs and novel CDSs, the remaining TSSs were used to identify antisense transcripts for all genes of *Actinoplanes* sp. SE50/110.

Only TSSs were examined which did not overlap a coding region on the sense strand. It was found that 99 of 845 TSSs either overlapped coding regions on the complementary strand (n=73) or were located within the promotor region of an antisense gene (n=26). **Table A.2** lists the affected genes.

3.4.3. Expression analysis of *Actinoplanes* sp. SE50/110 grown in three different cultivation media

Samples from all cultivation flasks for each of the three media conditions – 4× Mal-MM, 4× Mal-MM-TE, and 2× Glc-CM – were pooled prior to mRNA isolation using TRIzol (Life Technologies) and the RNeasy Mini Kit (QIAGEN). The cDNA library preparation was carried out with the help of the TruSeq RNA sample prep Kit (Illumina) as described in the Materials and Methods **Section 2.6.3**. Each of the three prepared libraries was then loaded on one lane and sequenced on an Illumina GA IIx platform, yielding between 11.7 and 5.9 million reads of 26 bp length. Subsequent strict quality filtering and exclusion of reads with ambiguous matching positions reduced the numbers to 1,843,987 (Mal-MM), 1,678,224 (Mal-MM-TE) and 1,200,691 (Glc-CM) reads of high quality (**Fig. 3.29**). The differences between library readouts were then normalized with the `DESeq` software [Anders & Huber, 2010]. A proportion of 24-34% normalized reads overlapped previously annotated CDS on the *Actinoplanes* sp. SE50/110 genome and were used for further analysis of DE genes and gene clusters.

Due to the fact that all reads were excluded from the analysis that mapped to multiple genomic loci, only few reads overlapped the ribosomal RNA genes because of their six-fold occurrence in the genome [Mehling *et al.*, 1995a, Schwientek *et al.*, 2012]. In contrast, a great amount of reads was mapped to transfer RNA genes, which also varied to a great extent between the conditions (9-33%). Because low levels of tRNAs can indicate cellular stress situations like starvation or oxidative stress [Haiser *et al.*, 2008], the differences between the conditions might reflect the cells' biosynthetic activity levels and growth phases, respectively. Consequently, the Mal-MM condition may not have reached stationary phase at harvest time as opposed to Mal-MM-TE and Glc-CM conditions, where significantly less proportions of tRNA reads were sequenced.

Overall, 11-16% of the annotated genes were found to be unexpressed, with not a single read matching to their CDS. Moreover, most of the remaining genes are very weakly expressed given by the fact that 60-65% of the lowest abundant genes are covered by only 5% of the reads. On the other hand, the 5% (413 genes) with the highest expression rate account for 54-63% of all available reads. This bias towards very few but highly expressed genes is consistent with the literature [Łabaj *et al.*, 2011] and poses putative targets for future knock-out or deletion experiments, aiming at decreasing energy and nutrient expenditure for unnecessary cell activities as described before for other actinomycetes [Baltz, 2011].

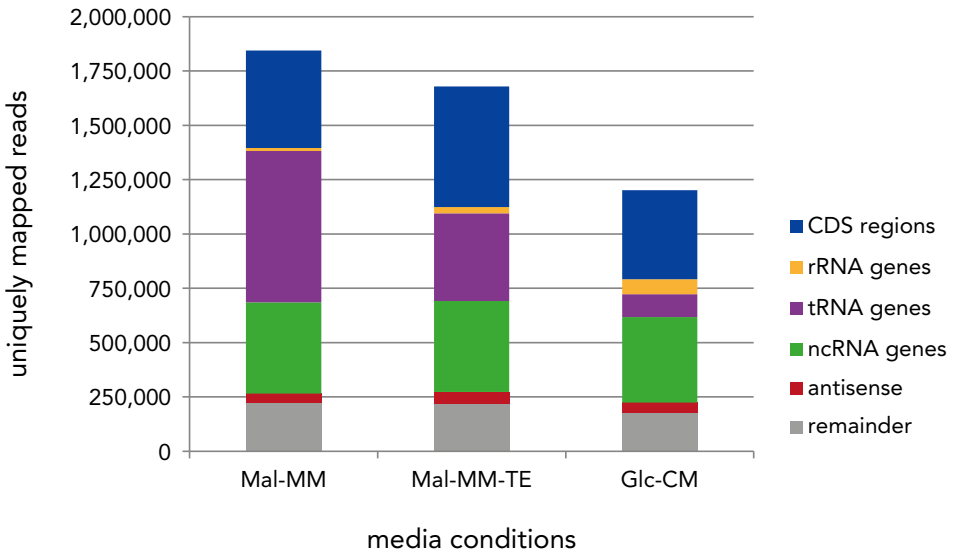

Figure 3.29.: Composition of sequenced reads from *Actinoplanes* sp. SE50/110 cultured in three different growth media. The raw read counts are shown prior to normalization. The bars represent one lane of an Illumina GA II*x* run for each of the three cultivation conditions. Each bar shows the number of reads that overlapped coding sequence regions (blue), ribosomal RNA genes (orange), transfer RNA genes (purple), and non-coding RNA genes (green). Additionally, the amount of reads that act as putative antisense transcripts are depicted (red) along with the number of reads mapping to none of the before mentioned genomic features (gray). The media conditions are abbreviated as maltose containing minimal medium (Mal-MM), Mal-MM with trace elements (Mal-MM-TE) and glucose containing complex medium (Glc-CM).

Highly expressed genes of *Actinoplanes* sp. SE50/110 cultivated in three different growth media play a role in transcriptional and translational processes

The identification of genes with high expression levels provides important insight into the cellular processes of *Actinoplanes* sp. SE50/110. It is known that the most abundant proteins in bacterial cells are usually associated with proliferation and maintenance functions [Ghaemmaghami *et al.*, 2003]. Consequently, it is assumed that their mRNA levels are also comparably high, although a clear dependency between gene expression and translation can be clouded by posttranscriptional regulation, e.g. through small ncRNAs [Massé *et al.*, 2003, Gottesman, 2005]. The most highly expressed genes were determined by the sum of their normalized read counts over all conditions.

An analysis of the 431 (5%) strongest expressed genes revealed a strong emphasis on ribsosomal associated proteins (57 genes) and proteins involved in transcriptional processes (51 genes). Furthermore, protein modification mechanisms (29 genes) and signal transduction (21 genes) together with energy production (22 genes) and carbo-

hydrate metabolism (17 genes) account for the major categories associated with the remaining functional annotated genes. Notably, despite their high expression rate, for 200 of the genes (46%) no reliable annotation was available.

The top twenty genes were examined in more detail between all cultivation conditions (**Tab. 3.12**). Interestingly, three of them encode conserved membrane proteins of unknown function, including the most highly expressed gene *acpl3986*. Of all ribosomal proteins, RpmI (Acpl6445) and RpmG (Acpl736) show the strongest expression, whereas the most prominent transcriptional regulator Acpl8038 has striking protein sequence similarity to the CarD family of transcriptional regulators, especially to its homologue from *Salinispora tropica* (98% identity). It was shown that CarD regulates light-induced carotenogenesis and fruiting body formation in response to nutrient limitation in *Myxococcus xanthus* [Padmanabhan *et al.*, 2001, Cayuela *et al.*, 2003]. However, in view of its high expression level in *Actinoplanes* sp. SE50/110, it is likely that CarD also acts as an architectural factor that aids in the assembly of protein complexes that are essential for DNA transcription, replication or repair as proposed earlier [Elías-Arnanz *et al.*, 2010]. Another highly expressed gene, *cgt* (*acpl5091*), might be of special interest regarding acarbose production because of its predicted function as secreted starch binding enzyme. While the small protein (149 amino acids) merely consists of two starch binding domains, it exhibits high similarity to the C-terminal domain of cyclodextrin glycosyltransferases and is therefore likely to be involved in carbohydrate utilization. In this regard, Cgt may enhance and support extracellular carbohydrate degrading enzymes, such as the pullulanase PulA and, potentially, the alpha-amylases AcbE and AcbZ encoded within the acarbose gene cluster [Wehmeier & Piepersberg, 2004]. This is in line with recent proteome studies which clearly identified Cgt, PulA, AcbE, and AcbZ in the exoproteome of *Actinoplanes* sp. SE50/110 cultures [Wendler, p.c.]. The high expression of *cgt* in the acarbose production media in contrast to almost no expression in the Glc-CM medium furthermore indicates an induction through maltose, similar to the induction of the acarbose gene cluster. Interestingly, the maltose importer operon *malEFG*, which might also contain the adjacent downstream *pulA* gene, is clearly expressed. However, no significant DE could be observed between the three cultivation conditions. This holds also true for the adjacent transcriptional regulator gene of the maltose importer operon, *malR*, located upstream of *malE* in the reverse orientation. MalR was previously thought to be a promising candidate for the regulation of the acarbose gene cluster.

Supplementation of trace elements induces the expression of genes involved in oxidative stress in *Actinoplanes* sp. SE50/110

On the cell physiological level, the addition of trace elements to Mal-MM lead to an increase of cell dry weight and total acarbose production in the Mal-MM-TE condition (**Fig. 3.18**). In order to investigate the underlying changes in gene expression, the most DE genes between both conditions were analyzed by means of the `DESeq` software [Anders & Huber, 2010]. The analysis revealed 70 ($\sim$1%) significantly ($p < 0.05$)

Table 3.12.: The twenty highest expressed genes in all three RNA-seq cultivation conditions.

Gene	Gene symbol	Mal-MM RPKM	Mal-MM-TE RPKM	Glc-CM RPKM	Product
acpl3986		35077	137123	61605	predicted membrane protein
acpl6445	*rpmI*	36909	19230	32229	ribosomal protein L35
acpl5091	*cgt*	54915	23454	489	starch binding domain containing secreted protein
acpl8038	*carD*	16935	17280	24153	transcriptional regulator CarD family
acpl2698		20691	17935	12550	hypothetical protein
acpl763	*rpmG*	15721	16520	12048	ribosomal protein L33
acpl4008		6070	25245	11556	predicted membrane protein
acpl7610		15392	19900	7517	hypothetical protein
acpl3976		5675	12836	16511	hypothetical
acpl4235		6670	7690	20514	hypothetical membrane protein
acpl7623	*cspA*	15392	9350	8484	Cold shock protein CspA
acpl7205	*rpsO*	12223	8352	11896	ribosomal protein S15
acpl7608		10788	21220	424	hypothetical protein
acpl7115		8825	10910	10663	hypothetical protein
acpl6562	*cspD*	10458	7496	9790	Cold shock domain protein CspD
acpl2290	*sdpR*	16640	9014	1420	transcriptional repressor SdpR
acpl7465	*rpmE*	9373	7676	9202	ribosomal protein L31
acpl1394	*rpsT*	10442	6975	8533	ribosomal protein S20
acpl1583		12205	8171	5009	hypothetical protein
acpl3833		12113	10707	1268	predicted integral membrane protein

DE genes (**Fig. 3.30**). Of these, six genes which are organized in two adjacent operons are most prominently up-regulated in Mal-MM-TE (**Fig. 3.31**).

The first operon encodes a transcriptional regulator of the CopY family as well as a zinc-dependent protease (Acpl3030). CopY is the transcriptional repressor of the *copYZAB* operon, which encodes proteins for the regulation of copper homeostasis in *Enterococcus hirae* [Magnani & Solioz, 2005] and other Gram-positive bacteria [García-Castellanos *et al.*, 2004]. In *E. hirae*, the repressor activity of CopY is deactivated by CopZ-mediated copper donation at increased cellular copper levels. At high levels however, CopZ is subject to proteolytic degradation by an unknown protease, as its excess is believed to be toxic for the cells [Lu *et al.*, 2003]. In this view, it is tempting to speculate that Acpl3030 undertakes this proteolytic function in *Actinoplanes* sp. SE50/110. In contrast to *E. hirae*, the *copYZAB* homologues in *Actinoplanes* sp. SE50/110 are not arranged in a consecutive operon and may therefore be regulated in more complex ways. Accordingly, the homologue to *copA*, encoding a putative copper import ATPase, is up-regulated four-fold, whereas its counterpart *copB*, encoding a putative copper export ATPase, is slightly down-regulated in Mal-MM-TE. These findings suggest that *Actinoplanes* sp. SE50/110 harbors a similar but more complex regulated copper homeostasis system than *E. hirae*. In addition, *Actinoplanes* responds positively to the supplied amount of copper, which acts as an important cofactor in many enzymes like lysyl oxidases, tyrosinases, Cu/Zn superoxide dismutases, and cytochrome c oxidases [Lu *et al.*, 2003].

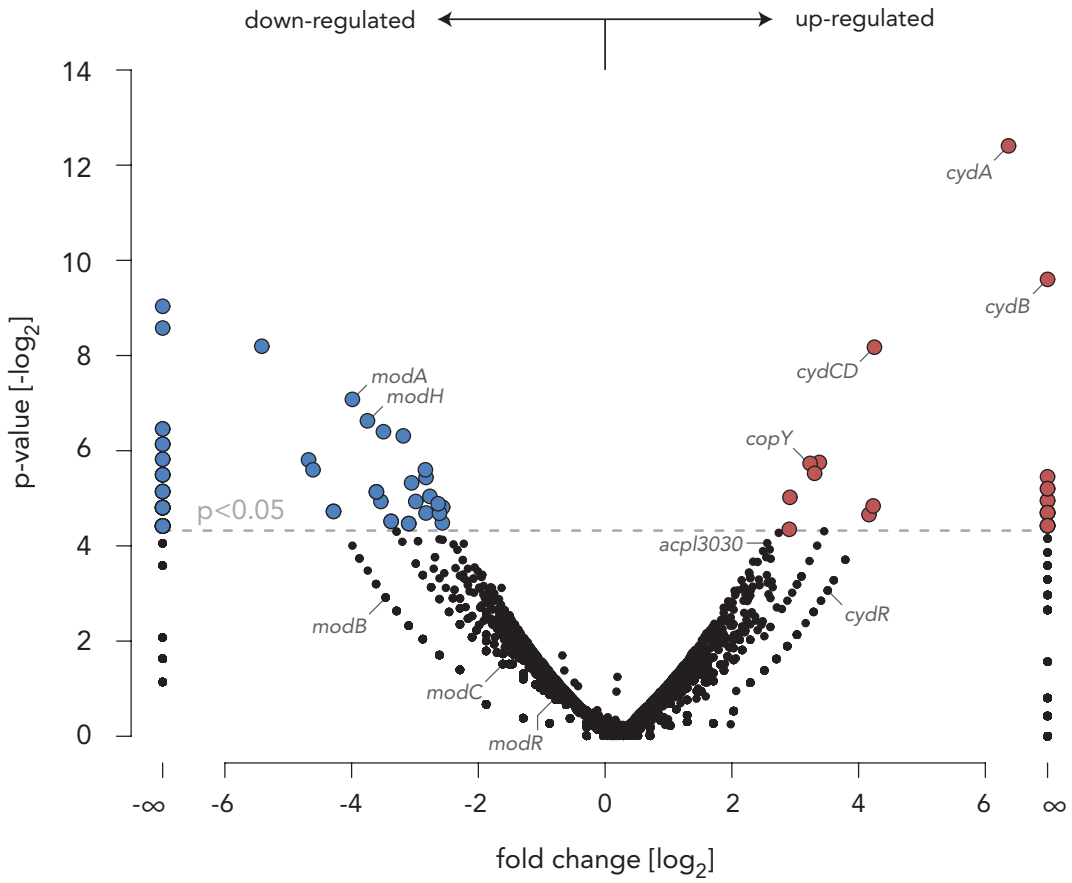

Figure 3.30.: Differential gene expression of *Actinoplanes* sp. SE50/110 cultivated in minimal medium (Mal-MM) and in minimal medium supplemented with trace elements (Mal-MM-TE). The volcano plot shows the fold change of read counts for all genes in the Mal-MM-TE condition with respect to their read counts in the Mal-MM condition. The genes above the significance threshold ($p < 0.05$) are marked in blue (down-regulated in Mal-MM-TE) and red (up-regulated in Mal-MM-TE). The genes which are discussed in the text are shown near their corresponding spot. Genes with zero reads in one of the conditions cause an infinite fold change, these genes are depicted at the outermost positions in the diagram.

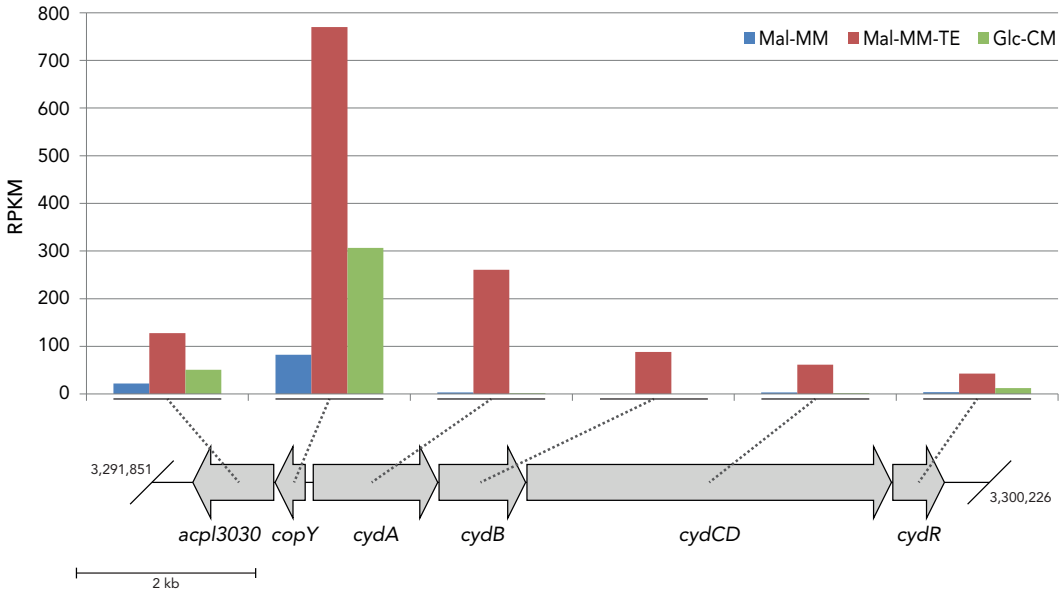

Figure 3.31.: The two most prominently up-regulated gene clusters of *Actinoplanes* sp. SE50/110 when comparing Mal-MM against Mal-MM-TE conditions. The left operon encodes the transcriptional regulator CopY and the zinc dependent protease Acpl3030, whereas the right operon harbors four genes that encode a cytochrome *bd* oxidase complex.

The second operon encodes a complete cytochrome *bd* oxidase gene cluster *cydABCD* with an additional transcriptional regulator CydR. Cytochrome *bd* is a widespread oxidase found in aerobic bacteria [Kranz & Gennis, 1985] and some archaea [Mathias, 1995], where it is responsible for the detoxification of dioxygen through reduction to water as terminal part of the respiratory chain. The cytochrome complex consists of two subunits, encoded by *cydA* and *cydB*, which form an integral membrane heterodimer. The *cydABCD* operon usually encodes two additional genes *cydC* and *cydD*, encoding two ABC-type transporter proteins required for the assembly of the complex [Das *et al.*, 2005]. In *Actinoplanes* sp. SE50/110 however, these are fused to a single gene *cydCD* of about 4 kb length as determined by the using searching the CDD [Marchler-Bauer *et al.*, 2011b] and comparative analysis between *cydCD* and its homologues from *Bacillus subtilis* [Winstedt *et al.*, 1998]. It is known from several bacteria such as *B. subtilis* [Winstedt *et al.*, 1998] and *Streptomyces coelicolor* [Brekasis & Paget, 2003] that *cydABCD* expression is induced under oxygen limiting conditions. As the average *cydABCD* up-regulation was found to be 60-fold in comparison to the Mal-MM condition (**Fig. 3.31**), it is likely that the culture reached oxygen limitation as a result of increasing cell density, which was triggered by the addition of growth promoting trace elements in the Mal-MM-TE medium.

The additional transcriptional regulator CydR shows highest sequence similarity to the TetR family of transcriptional repressors, which generally control gene expres-

sion of products involved in multidrug resistance, biosynthesis of antibiotics, osmotic stress and others [Ramos *et al.*, 2005]. In *Actinoplanes* sp. SE50/110, the CydR 5'-coding sequence shares a 29 bp overlap with the 3'-end of *cydCD*, which suggests its participation in the *cyd* operon. In spite of this, CydR exhibits only poor sequence similarity to the known regulators ArcA, FNR [Patschkowski *et al.*, 2000], YdiH/Rex [Brekasis & Paget, 2003, Schau *et al.*, 2004], CcpA or ResD [Puri-Taneja *et al.*, 2007] controlling the expression of *cydABCD* in other species. As a result, the function of CydR remains to be determined.

Another gene cluster, *modHABCR*, showing clear DE between Mal-MM and Mal-MM-TE is related to molybdenum (Mo) uptake (**Fig. 3.32**), which is in line with Mo being one of the supplied trace elements. Molybdate, which is the bioavailable form of Mo, plays an essential role in microbial metabolism because of its requirement as cofactor in important enzymes such as nitrate reductase or formate dehydrogenase [Maupin-Furlow *et al.*, 1995]. The *mod* cluster of *Actinoplanes* sp. SE50/110 exhibits structural similarity to the molybdate import system from *E. coli* (*modABCD*) in the sense, that the genes *modABC* share common functional annotation between both organisms. Analysis of the proteins in *E. coli* showed that ModA functions as a molybdate-specific periplasmic binding protein, ModB as an integral membrane channel-forming protein, and ModC as an ATP-binding energizer protein [Grunden *et al.*, 1999]. In contrast to *E. coli* however, *Actinoplanes* sp. SE50/110 lacks the terminal *modD* gene, which is replaced by a putative transcriptional regulator named ModR. In addition, another protein designated ModH, which contains both, a helix-turn-helix DNA binding domain and a molybdenum binding domain, precedes the cluster. Although the sequence similarity between ModH and the transcriptional repressor ModE of the *mod* operon in *E. coli* is very weak, the domain structure seems to be conserved, indicating a possible regulatory function for ModH.

The observed gene expression of the mod cluster in *Actinoplanes* sp. SE50/110 is very low in Mal-MM-TE compared to Mal-MM, which is consistent with findings from *E. coli* where ModE repressed *modABCD* expression in the presence of increased intracellular molybdate levels [Grunden *et al.*, 1999]. Consequently, a high *mod* expression hints to molybdate limitation in Mal-MM and Glc-CM media which may also be a reason for the lower cell dry weights measured of these conditions (**Fig. 3.18**).

Genes involved in metal metabolism are differentially expressed between the three growth media of *Actinoplanes* sp. SE50/110

The comparison of Glc-CM and Mal-MM gene expression levels permits the analysis of transcriptional changes that lead to the biosynthesis of acarbose in the Mal-MM condition. It has to be expected however, that these changes are accompanied or even clouded by effects resulting from differences in the media which do not directly impact the acarbose biosynthesis. As the Glc-CM medium on the one hand contains complex ingredients (yeast extract and peptone) and glucose as carbon source, whereas Mal-MM contains maltose and several defined trace elements in rather high quantities on

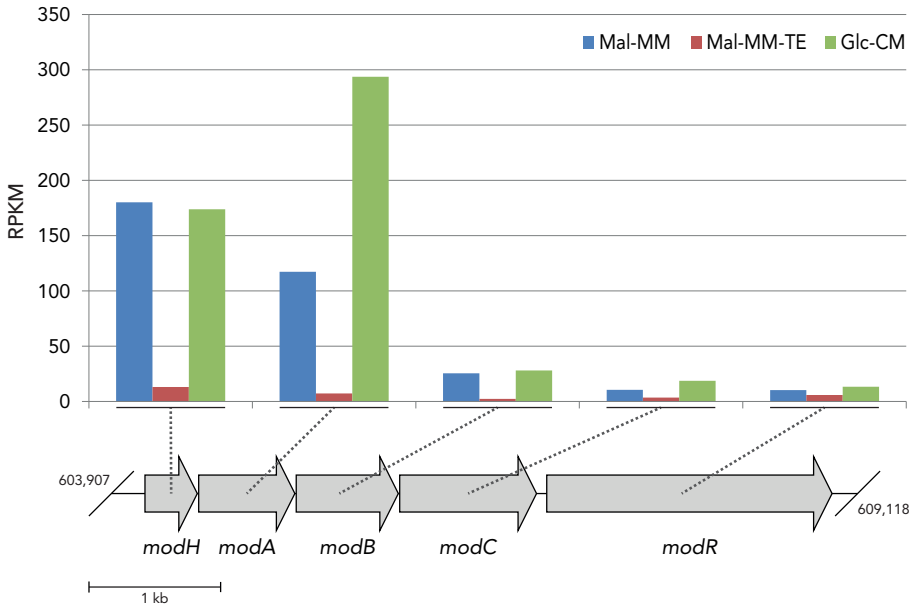

Figure 3.32.: The most prominently down-regulated gene cluster of *Actinoplanes* sp. SE50/110 when comparing Mal-MM against Mal-MM-TE conditions. The cluster harbors five genes that encode a molybdenum uptake system.

the other hand, a much more diverse expression pattern (**Fig. 3.33**) was observed, as opposed to Mal-MM and Mal-MM-TE (**Fig. 3.30**). Consequently, a total of 546 (∼6%) significantly ($p < 0.05$) DE genes were identified.

The most striking change in expression was found in the bacterioferritin BfrA and the bacterioferritin-associated ferredoxin Bfd. Bacterioferritin assembles to a 24mer cluster of roughly spherical shape which acts as an iron storage protein that regulates iron availability within the cell and may also be involved in iron detoxification and other processes [Carrondo, 2003]. While BfrA exhibits excellent sequence similarity of up to 89% to bacterioferritins of various *Streptomyces* strains, close homologues to Bfd are rarely found in public databases with the exception of *Stackebrandtia nassauensis* with 85% similarity [Munk *et al.*, 2009]. Bfd from *E. coli* is believed to participate in iron storage through intracellular iron transport, mobilization of bacterioferritin, or regulation [Garg *et al.*, 1996,Quail *et al.*, 1996]. It is remarkable that both genes are literally silent (zero read count) in the Mal-MM condition and strongly induced in the Glc-CM medium (**Fig. 3.34A**). This leads to the hypothesis that iron availability negatively regulates the transcription of *bfrA* and *bfd*, possibly through a transcriptional repressor like the ferric-uptake regulator Fur, known to occur in many species [Escolar *et al.*, 1999]. Fur is a general regulator of iron-dependent expression of more than 90 genes in *E. coli* and functions as a positive repressor, in the sense that it needs iron as co-repressor in order to bind to its target DNA-sequence and inhibit transcription [Andrews *et al.*, 2003]. In support of this supposition, a putative Fur binding sites was identified 92 bp upstream of the *bfd* start codon

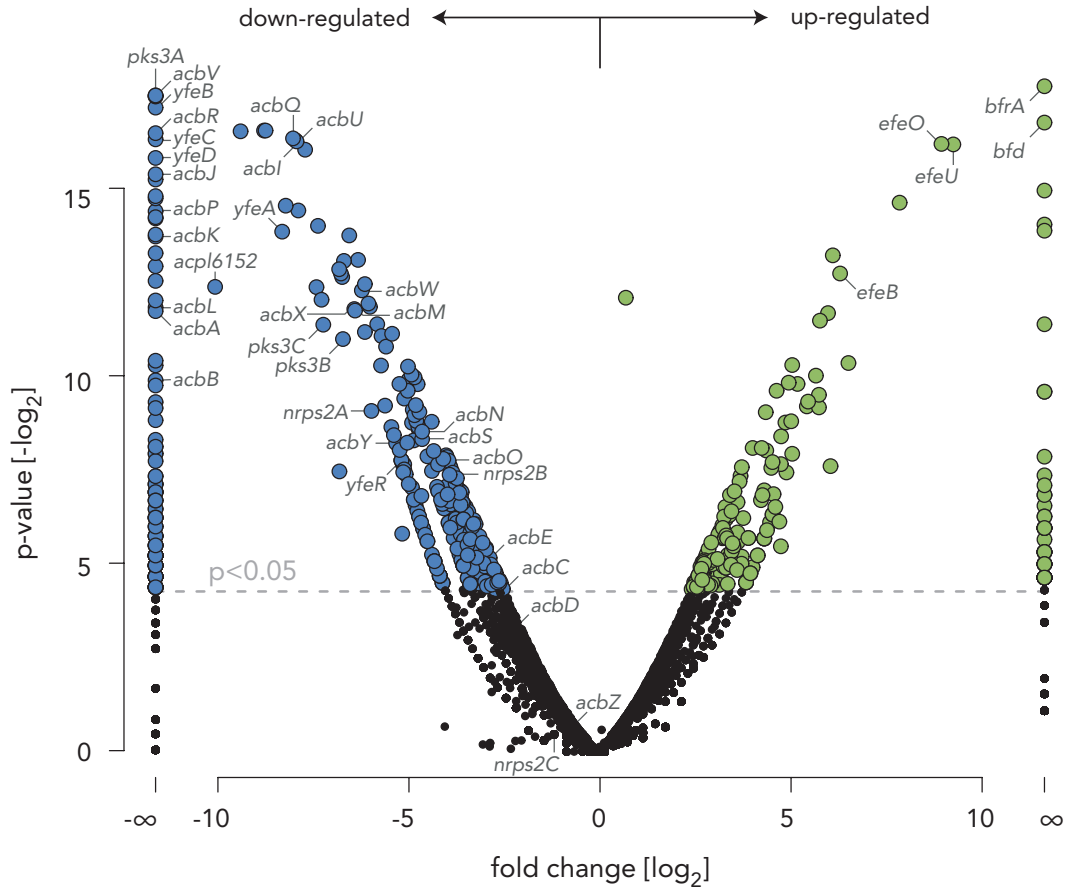

Figure 3.33.: Differential gene expression of *Actinoplanes* sp. SE50/110 cultivated in minimal medium (Mal-MM) and in complex medium (Glc-CM). The volcano plot shows the fold change of read counts for all genes in the Glc-CM condition with respect to their read counts in the Mal-MM condition. The genes above the significance threshold ($p < 0.05$) are marked in blue (down-regulated in Glc-CM) and green (up-regulated in Glc-CM). The genes which are discussed in the text are shown near their corresponding spot.

showing 12/19 identities to the Fur consensus sequence of *E. coli* [Escolar *et al.*, 1999].

A gene cluster with comparable DE ratios to *bfd*/*brfA* was found to encode three proteins, which resemble a relatively new class of iron importers [Debut *et al.*, 2006], best studied in *E. coli* (EfeUOB) and its homologue YwbLMN from *B. subtilis* [Ollinger *et al.*, 2006, Cao *et al.*, 2007]. It was shown for both organisms that their respective cluster is also Fur-regulated (repressed by iron), which suggests a similar regulation in *Actinoplanes* sp. SE50/110 based on the striking up-regulation of the cluster in the iron-limited Glc-CM condition (**Fig. 3.34B**). However, no clear Fur-binding site could be identified. While EfeU was shown to be an integral-membrane iron-permease, the exact functions of the other two proteins have not yet been completely unraveled [Rajasekaran *et al.*, 2010]. It was proven, however, that all components are necessary to form a functional iron importer in *E. coli* [Cao *et al.*, 2007]. The corresponding proteins from *Actinoplanes* sp. SE50/110 show highest sequence similarities of 72-86% to homologous clusters of the two *Micromonospora strains* L5 and ATCC 39149 and were designated EfeUOB according to their *E. coli* homologues.

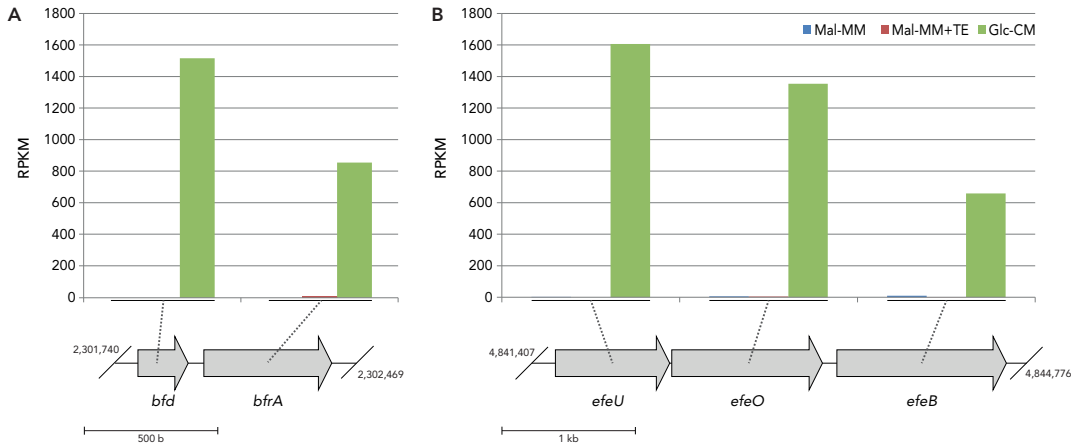

Figure 3.34.: The most prominently up-regulated gene cluster of *Actinoplanes* sp. SE50/110 when comparing Mal-MM against Glc-CM conditions. (**A**) The regulation of the bacterioferritin gene cluster and (**B**) the expression of the iron importer operon *efeUOB*

A third, highly DE genomic locus, was identified because of its extreme repression in the Glc-CM condition (**Fig. 3.35**). The cluster consists of five genes, resembling the *yfeABCD* operon of *Yersinia pestis*, which encodes an ABC metal transport system [Bearden *et al.*, 1998]. An additional gene *yfeR*, which encodes a predicted transcriptional regulator, overlaps 4 bp with the 5'-end of *yfeD*. The complete operon is silent in the Glc-CM condition and highly up-regulated in both Mal-MM media. This is astonishing because *yfe* as well as a homologous cluster from *Actinobacillus pleuropneumoniae* were shown to be up-regulated under iron restriction [Bearden *et al.*, 1998, Deslandes *et al.*, 2007], which would be in line with the identifica-

tion of a Fur binding site (14/19 identities) upstream of the *yfeABCDR* operon in *Actinoplanes* sp. SE50/110. While this contradiction could be explained with a rare case of Fur-induced expression [Andrews *et al.*, 2003], the participation of another regulator, possibly *yfeR*, is also conceivable.

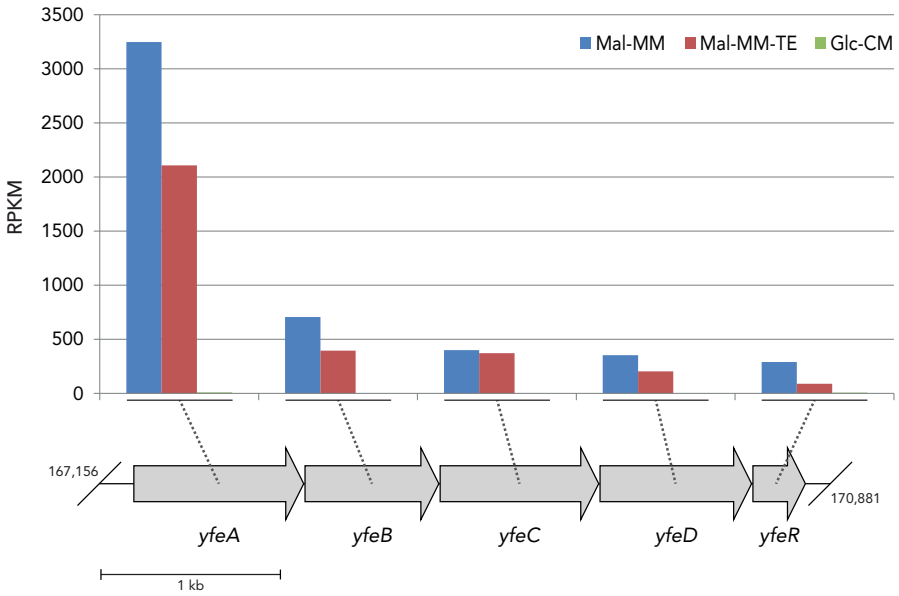

Figure 3.35.: The most prominently down-regulated gene cluster of *Actinoplanes* sp. SE50/110 when comparing Mal-MM against Glc-CM conditions. The cluster harbors five genes that encode the ABC-type iron importer operon *yfeABCD* and its putative regulator YfeR.

However, from a cell physiological point of view, the observed expression pattern of iron-transport related genes are likely to reflect an oversaturation of iron in the Mal-MM cells, which necessitates the observed repression of iron importers and the induction of the putative export system YfeABCD(R). It can therefore be concluded that the Glc-CM medium exposes *Actinoplanes* sp. SE50/110 to iron limiting conditions whereas both Mal-MM media contain sufficient or even an excess of iron. In this context, it should be noted that the above mentioned oxidative stress response could also be a consequence of iron excess, as high intracellular levels of iron are toxic and lead to oxidative stress through the generation of reactive oxygen species under aerobic conditions [Groves *et al.*, 2010, Cornelis *et al.*, 2011].

A secondary metabolite biosynthesis gene cluster of *Actinoplanes* sp. SE50/110 is highly expressed in the acarbose production media

Recently, four putative antibiotic gene clusters were identified in the *Actinoplanes* sp. SE50/110 genome sequence [Schwientek *et al.*, 2012]. One of them, cACPL_4 consists of three NRPS, three PKS and several accessory proteins including a possible

product exporter complex, which constitute a hybrid NRPS/PKS cluster. Interestingly, this locus was found to be highly up-regulated in both Mal-MM media, whereas a comparable expression in the Glc-CM medium was only found for one of the NRPS genes *nrps2C* and the adjacent ABC-type multidrug transporter genes *acpl6159-6161*, located at the periphery of the cluster (**Fig. 3.36**).

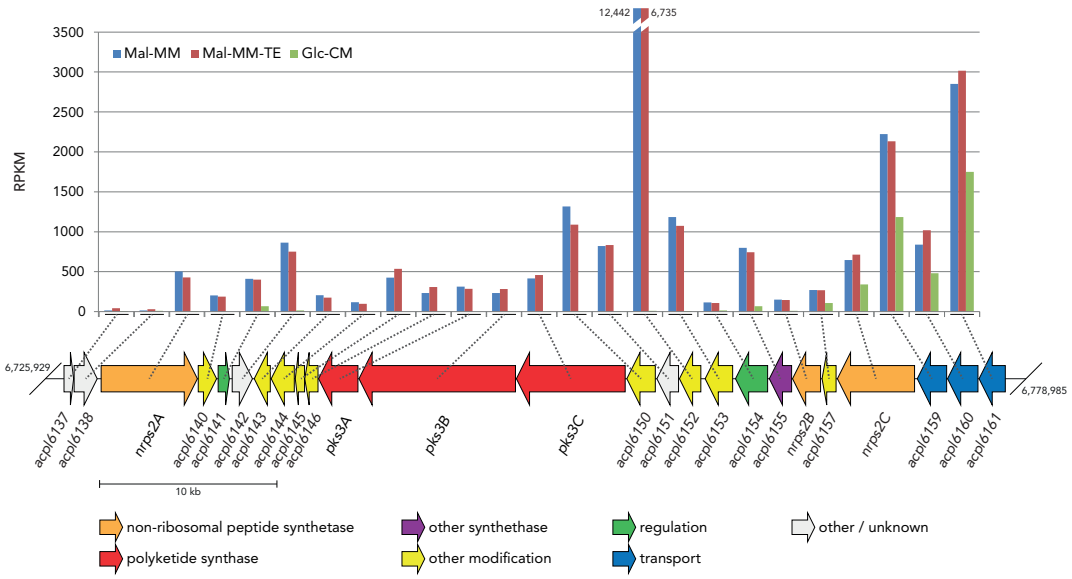

Figure 3.36.: The significantly down-regulated hybrid NRPS/PKS cluster of *Actinoplanes* sp. SE50/110 when comparing Mal-MM against Glc-CM conditions. Most genes of the putative antibiotic biosynthesis gene cluster cACPL_4 are strongly down-regulated in the complex medium, whereas their expression is comparably high in both minimal media. The strong expression of the gene *acpl6152*, encoding a putative tryptophan halogenase, is particularly noticeable because this enzyme catalyzes the first step in the biosynthesis of pyrrolnitrin, an antibiotic with anti-fungal activity known from *Pseudomonas pyrrocinia*.

These findings are of tripartite implication. First, the genes *acpl6158-6161* seem to be regulated in a common way, which obviously differs from the regulation of the remaining cluster, as their high expression rates were also found in the Glc-CM condition. This suggests a rather constitutive function for the NRPS and its adjacent multidrug transporter, which might not be linked to the remainder of the cluster as previously anticipated. Second, the expression of the complete cluster is induced equally well in both Mal-MM conditions, ruling out a possible stimulation through trace elements only supplied in the Mal-MM-TE medium. It remains to be determined which of the Mal-MM medium components caused the induction, however. Third and most importantly, the high expression of the cluster implies a significant investment of compounds and energy from the cell, which might rather be channeled towards an increase in acarbose production. In terms of mRNA measurements, 1.7% of all filtered reads that mapped to genes were contained in this cluster. It should also be noted

that the first two genes *acpl6137* and *acpl6318* exhibit very low expression and are therefore likely not to be involved in the NRPS/PKS hybrid cluster.

The acarbose biosynthetic gene cluster of *Actinoplanes* sp. SE50/110 is highly expressed in maltose containing media and silent in the glucose containing medium

The transcriptional analysis of the acarbose gene cluster confirms previous experiments that found glucose, in the absence of maltose, to cause stalling of the acarbose production, presumably through a down-regulation of the acarbose gene expression [Brunkhorst & Schneider, 2005, Wang *et al.*, 2011b]. It is therefore not surprising that the *acb* gene cluster was scarcely expressed in the Glc-CM condition (**Fig. 3.37**). Merely the genes of the extracellular alpha-amylases AcbZ and AcbD were expressed at a moderate level, accompanied by the C7-cyclitol cyclase AcbC and the Cyclitol-7-phosphate 2-epimerase AcbO, which catalyze the first and third step in the biosynthesis of acarbose (**Fig. 1.5**).

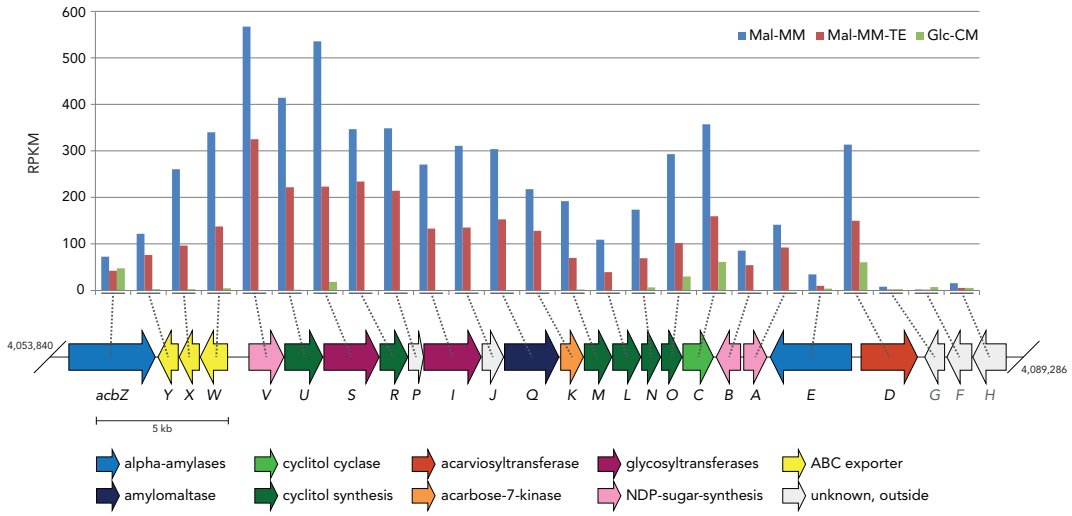

Figure 3.37.: The regulation of the acarbose biosynthetic gene cluster of *Actinoplanes* sp. SE50/110 with the adjacent galactose importer system *acbHFG*. Interestingly, the expression of the genes is approximately twice as high in Mal-MM compared to Mal-MM-TE, whereas almost no expression was detected in the complex medium Glc-CM.

Both maltose containing media exhibit a clear induction of the *acb* gene cluster. Interestingly, the expression of the cluster is approximately two-fold down-regulated when trace elements were added. Although the gene's differential expressions are not significant between Mal-MM and Mal-MM-TE, this regulation might indicate a positive effect of the supplied trace elements, as the acarbose production yields are about the same for both conditions when normalized to cell dry weights (44 mg as opposed to 46 mg per gram of CDW). Hence, the down-regulation did not alter acarbose

productivity. One possible explanation for this finding may be derived from the aforementioned lack of needed trace elements in Mal-MM, which could have hindered the correct assembly or functioning of certain copper, zinc, molybdenum or manganese dependent enzymes, which necessitated a higher expression to yield the same amount of functional enzymes. However, further research has to be conducted to answer this question, as e.g. a different growth phase of the conditions could certainly result in a similar expression pattern.

Recent crystallographic analysis of the extracellular binding protein AcbH raised evidence for the putative acarbose importer AcbGFH to bind galactose, rather than acarbose [Licht *et al.*, 2011]. Given the distinct low expression pattern of *acbGFH* (**Fig. 3.37**), the analysis clearly supports these findings, as an acarbose importer would expected to be expressed at an elevated level in order to comply with the carbophore hypothesis [Wehmeier & Piepersberg, 2004].

Chapter 4.
4 Discussion

In this chapter, major findings of the previously reported results are critically examined with regard to the current literature. In particular, novel insights are emphasized that enhance the current knowledge about the acarbose metabolism and putative ways to enhance its production in *Actinoplanes* sp. SE50/110. Furthermore, outcomes of general scientific interest are highlighted and based thereon improved procedures for similar research projects are elaborated. Finally, aspects of the new RNA-sequencing technology and its bioinformatic analysis are discussed.

4.1. Establishment of the complete *Actinoplanes* sp. SE50/110 genome sequence

Despite the advances in second generation sequencing, such as the tremendous throughput and low sequencing costs per base, finishing of genome sequences becomes decreasingly common [Chain *et al.*, 2009]. The reasons for this trend are mainly attributed to the additional investment of significant amounts of time and money, which are usually out of proportion to the additionally gained sequence information. While complete genome sequencing of procaryotes with moderate GC-content seems to be rather straightforward, second generation sequencing of high-GC organisms still poses difficulties [Dohm *et al.*, 2008]. The present study exemplified this fact especially well, since the initial sequencings using standard chemistry and PE protocol yielded rather poor results. This, however, also shows what might be missed by generally leaving genomes in draft status. In the case of *Actinoplanes* sp. SE50/110, a total of 906 kb, accounting for roughly 10% of the complete genome size, were missed in the initial runs. Furthermore, the high level of fragmentation renders draft genomes unsuitable for a variety of analyses such as core-genome based comparative phylogenies and *in silico* pathway modeling, because decisive gene information might be missing. In this regard, the disrupted sequence information, as exemplified in **Section 3.1** for the acarbose gene cluster, called on a detailed analysis of the underlying causes.

Importantly, the described outcome thereof – the identification of high-GC islands and secondary structure formation within gap regions – might be of outstanding interest to other sequencing projects working with high-GC organisms and second generation technologies. Without doubt, the enhancements of the sequencing process which were derived from these findings lead to a significant improvement of the *Actinoplanes* sp. SE50/110 genome sequence with only 1.6% of missing sequence information and a strongly reduced genome fragmentation. This was achieved by using longer read length, which improved the decomposition of shorter repetitive elements, and trehalose as a chemical additive in the emulsion PCR step of the library preparation protocol, which reduced secondary-structure formation of single stranded DNA during amplification. On that account, the demonstrated feasibility and successful application for high-GC genome sequencing using these enhancements on the Genome Sequencer FLX platform has great potential for improving the quality of modern actinomycetes sequencing projects, as its application is simple, cheap, fast and effective.

While the results of the initial PE runs were far from optimal, their true value became evident during the combined assembly of the PE and WGS runs. Without PE information, the order and orientation of the 600 contigs would have been speculative, which would have complicated the finishing procedure to a great extent. By using the PE data, 421 contigs could already be positioned and primer-pairs of adjacent contigs were selected for PCR amplification and subsequent sequencing of the gap regions. It should therefore be noted that a combination of PE and WGS approaches is certainly beneficial for sequencing more complex genomes such as that of *Actinoplanes* sp. SE50/110. Nonetheless, with increasing read length and throughput of the technologies, it might be sufficient to use only long PE protocols with anti self-annealing substances in future projects in order to yield similar results from just a single run.

4.2. Annotation of the *Actinoplanes* sp. SE50/110 genome sequence

The genome annotation is a critical process that has to be performed with special care, since many analyses build upon its findings. Consequently several different gene finders were tested for their performance during the annotation of the *Actinoplanes* sp. SE50/110 genome. Of these, `Prodigal` performed best in terms of the resulting genome coding density (see **Table 3.6**), presumably because of its adaptivity to genomic GC-contents [Hyatt *et al.*, 2010]. Furthermore, it is long known from other high-GC genera such as *Streptomyces* that the third base of a codon exhibits a higher probability of being a guanine or cytosine if the codon is located in-frame of a coding sequence [Bibb *et al.*, 1984]. Manual inspection of the acarbose biosynthesis genes using the `FramePlot` software [Ishikawa & Hotta, 1999] confirmed this to be true also for *Actinoplanes* sp. SE50/110 and presumably all *Actinoplanes* species (see also **Section 3.3.1**). As the `Prodigal` gene finding algorithm makes heavy use of this frameplot-analysis, it is not surprising to produce good results for *Actinoplanes* sp. SE50/110 and might also be considered as standard gene finder for the annotation of other high-GC genomes in the future.

4.3. New insights related to the acarbose metabolism

4.3.1. Acarbose re-import after exclusion of *acbHFG*

The carbophore model attributes acarbose a second function as transport vehicle for oligosugars beside its main function as an inhibitor of competitor's α-glucosidases [Wehmeier & Piepersberg, 2004, Brunkhorst *et al.*, 2005]. Therfore, the model demands an uptake system for *loaded* acarbose, which was believed to be the *acbHFG* operon. However, in contrast to previous findings [Brunkhorst *et al.*, 2005], the extracellular binding protein AcbH, encoded within the *acbHFG* operon (**Fig. 3.15**), was recently shown to exhibit high affinity to galactose [Licht *et al.*, 2011] instead of acarbose or its homologues as was previously anticipated. This implies that *acbHFG* does not directly belong to the acarbose cluster anymore as was proposed by the carbophore hypothesis [Wehmeier & Piepersberg, 2004, Piepersberg *et al.*, 2002]. In order to search the *Actinoplanes* sp. SE50/110 genome for a new acarbose importer candidate, the *gacGFH* operon of a second acarbose gene cluster identified in *Streptomyces glaucescens* GLA.O has been used as query [Rockser & Wehmeier, 2009]. GacH was recently shown to recognize longer acarbose homologues but exhibits only low affinity to acarbose [Vahedi-Faridi *et al.*, 2010]. However, the search revealed rather weak similarities towards the best hit operon *acpl5404-acpl5406* with GacH showing 26% identity to its homologue Acpl5404. A consecutive search of the extracellular maltose binding protein MalE from *Salmonella typhimurium*, which has been shown to exhibit high affinity to acarbose [Vahedi-Faridi *et al.*, 2010], revealed 32% identity to its MalE homologue in *Actinoplanes* sp. SE50/110. Despite the low sequence similarities, these findings suggest that acarbose or its homologues are either imported by one or both of the above mentioned importers, or that the extracellular binding protein exhibits a distinct amino acid sequence in *Actinoplanes* sp. SE50/110 and can therefore not be identified by sequence comparison alone.

4.3.2. Putative formation of component C by trehalose synthases

Component C is structurally highly similar to acarbose and, thus, both compounds are difficult to separate in the industrial downstream processing of the fermentation broth [Wehmeier & Piepersberg, 2004]. In order to overcome this expensive and time consuming separation step, first attempts have been made to identify the reactions and enzymes that lead to the formation of component C. In this regard, it was demonstrated that component C is not formed by the acarviosyltransferase AcbD or any other extracellular enzyme as opposed to acarbose and its other homologues [Hemker *et al.*, 2001]. Rather, a unique way for component C synthesis was proposed involving trehalose synthases, which are believed to act directly on acarbose and its maltose moiety, respectively [Wehmeier & Piepersberg, 2004]. Following up on this idea, a study identified five genes encoding trehalose synthases, namely *otsA* (*acpl1307*) and *treS* (*acpl5330*) as well as the *treXYZ* operon (see **Table 3.7**) [Lee *et al.*, 2008]. The identified enzymes were overexpressed, purified and incubated with acarbose to test for their ability to convert acarbose into component C. The results

indicated that only TreY was able to perform the conversion reaction. It should be noted, however, that the reported results could not be observed when the experiment was repeated at the Wuppertal University [Dr. Wehmeier, p.c.].

With the complete genome sequence at hand, it was possible to identify three more *otsA* genes, a new *otsB* gene, and a second trehalose synthase (*treS*), which were tested in neither of the experiments. Hence, TreY might not be the (only) enzyme which is responsible for component C formation in *Actinoplanes* sp. SE50/110.

4.4. The actinomycete integrative and conjugative element pACPL

Perhaps the most interesting finding of all was the identification of the actinomycete integrative and conjugative element (AICE) pACPL. In regard to the aims of the whole project, pACPL could hold the key to genetic engineering of *Actinoplanes* sp. SE50/110 by serving as a transformational plasmid. Due to the nature of AICEs, which resemble the lifestyle of temperate phages to a certain degree, genetically modified plasmids would be able to integrate into the genome at specific attachment sites (*attB*), mostly located in tRNA genes. This might result in rapid production of transformants and, in contrast to autonomously replicating plasmids, the integration can be maintained without antibiotic selection and without reducing the yield of secondary metabolite production [Baltz & Hosted, 1996]. The feasibility of this approach has been shown for several *Strepomyces* AICEs including pSLP1 and pSAM2 (**Fig. 3.16**), which evolved to valuable tools for studying the biosynthesis of antibiotics in actinomycetes [Omer *et al.*, 1988, Smokvina *et al.*, 1990, Kuhstoss *et al.*, 1991]. In addition, a similar plasmid development strategy was used for *Micromonospora* spp. using the AICE pMR2 (**Fig. 3.16**) from *Micromonospora rosaria* [Hosted *et al.*, 2005].

In this context it should also be noted that several bacteriophages were described, which are capable of infecting *Actinoplanes* spp. [Jarling *et al.*, 2004a, Jarling *et al.*, 2004b]. In particular, a successful transformation system which is based on a bacteriophage was demonstrated for *Actinoplanes teichomyceticus*, the producer of the antibiotic teicoplanin [Ha *et al.*, 2008].

The newly identified AICE may also improve previous efforts in the analysis of heterologous promoters for the overexpression of the lipopeptide antibiotic friulimicin in *Actinoplanes friuliensis* [Wagner *et al.*, 2009].

Overall, these findings are of great interest, as they demonstrate the first native functional AICE for *Actinoplanes* spp. in general and imply the possibility of future genetic access to *Actinoplanes* sp. SE50/110 in order to perform targeted genetic modifications aiming at increased acarbose yields and elimination of component C formation.

4.5. The putative antibiotic gene clusters of *Actinoplanes* sp. SE50/110

In general, the four newly discovered secondary metabolite gene clusters broaden the knowledge of actinomycete NRPS and PKS biosynthesis clusters and represent

just the tip of the iceberg of the manifold biosynthetical capabilities – apart from the well-known acarbose production – that *Actinoplanes* sp. SE50/110 houses. It remains to be determined if all presented clusters are involved in industrially rewarding bioactive compound synthesis and how these clusters are regulated, because none of these metabolites were identified and isolated so far. These new gene clusters may also be used in conjunction with well-studied antibiotic operons, in order to synthesize completely new substances, as recently performed [Melançon & Liu, 2007, Oh *et al.*, 2007].

As stated in the Bayer HealthCare AG pharma reports from 1992 and 1993, the antibiotic thiazomycin was found in the cultivation broth of at least one of the acarbose overproducing strains (*Actinoplanes* sp. C445-P47) that were developed by mutagenesis experiments [Dr. Selber, p.c.]. Because all strains are based on *Actinoplanes* sp. SE50, it is very likely that also *Actinoplanes* sp. SE50/110 contains the corresponding biosynthesis gene cluster. Unfortunately, no thiazomycin cluster has been sequenced yet, which renders its identification within the *Actinoplanes* sp. SE50/110 genome sequences difficult. Based on the structurally similar antibiotic tyrocidine from *Bacillus brevis* however, it can be assumed that cACPL_1 (**Fig. 3.17A**) constitutes the responsible biosynthesis cluster. This is derived from the fact that tyrocidine is synthesized by three NRPSs which comprise ten modules for the incorporation of amino acids [Mootz & Marahiel, 1997]. Very similarly, cACPL_1 contains four NRPS genes comprising also ten modules. The other identified putative antibiotic gene clusters either lack NRPS genes or they do not contain enough modules to explain the size of the antibiotic thiazomycin, which is in fact a mixture of several derivatives similar to the antibiotics group of nocathiacins [Jayasuriya *et al.*, 2007].

Interestingly, it was found that thiazomycin is an extremely potent antibiotic against Gram-positive bacteria with potential clinical relevance e.g. in the treatment of infections with multi resistant *Staphylococcus aureus* strains [Singh *et al.*, 2007].

4.6. Transcriptome analyses of *Actinoplanes* sp. SE50/110

The transcriptome of *Actinoplanes* sp. SE50/110 was analyzed by conducting two distinct RNA-seq experiments for each of three cultivation conditions. In the case the of 5'-enriched cDNA libraries, the experiment was designed to identify TSSs within the *Actinoplanes* sp. SE50/110 genome sequence and indeed yielded more than 1,400 potential TSS from a pooled analysis of all three conditions. The other experiment was designed to provide quantitative expression values of full length transcripts that can be used for differential expression testing between the three conditions. The corresponding computational tests resulted in an impressive detection rate, given the fact, that no replicates were sequenced.

Overall, the advantages of RNA-seq over standard microarrays legitimate the higher costs and efforts in downstream analysis [Croucher & Thomson, 2010]. Performing RNA-seq experiments on organisms for which the genome sequence has not yet been established would certainly be even more rewarding and should be considered as an alternative to standard genomics approaches. This is due to the fact

that bacterial genomes exhibit a high coding density (**Tab. 3.6**), which might reveal a major part of the genome sequence after *de novo* transcriptome assembly of RNA-seq results. However, although costs for RNA-seq experiments are continuously reducing due to the steady increase of sequencing throughput, microarrays are still more cost efficient, especially if large numbers of experiments are to be carried out. Therefore, a good strategy might include an initial RNA-seq run, which optimizes the genome annotation, followed up by microarrays that were designed on the basis of the improved annotation.

In this regard it should also be noted that the high-GC content of *Actinoplanes* sp. SE50/110 might have caused similar self-annealing problems during the amplification step of the RNA-seq library preparation protocols as it did in the DNA library preparation (see **Section 3.1**). Although this was not explicitly analyzed in this work, comparisons with other RNA-seq runs of moderate-GC organisms that were sequenced on the identical device, resulted generally in more high quality reads.

4.6.1. Improvement of genome annotation by RNA-seq

Genome annotation improvement which is based on 5'-enriched RNA-seq data relies on the identification of TSSs for subsequent determination of genomic features. While this approach works well for CDS start site corrections and the detection of TSSs for all kinds of genomic features as described in **Section 3.4.2**, the actual transcript length might not be directly derived from these observations alone. This is especially cumbersome for ncRNAs, which, in contrast to mRNAs, also lack a stop codon that could be used to at least define the end of the functional sequence within the transcript. This limitation was circumvented in this and other studies by utilization of a whole transcript cDNA library, which was mainly created for the detection of differentially expressed genes [Wurtzel *et al.*, 2010]. In general, however, this might not be an optimal strategy, because the transcript length determination depends heavily on the sequencing depth of the run. In other words, it is impossible to determine the exact end of a transcript if it is not completely covered by sequenced reads. This obvious drawback has recently been recognized and lead to the development of a novel cDNA library preparation method for RNA-seq, termed *RNA paired-end tag* (RNA-PET) [Ruan & Ruan, 2012]. In principle, RNA-PET enriches both, the 5'- and the 3'-ends of full length cDNAs by discarding the central part of the transcripts (similar to Illumina's mate-pair library preparation [Van Nieuwerburgh *et al.*, 2011]). Additionally, the protocol assures that both ends of the transcript will be concatenated prior to sequencing, which yields PE information. In contrast to PE DNA-sequencing as described in **Section 3.1** however, the distances between the sequence pairs are variable in RNA-PET and match the lengths of the original transcripts. Thus, reference mapping of the PE data results in the determination of transcript length and additional information, such as alternative start- and stop-sites of a transcript [Ruan & Ruan, 2012]. RNA-PET is therefore better suited to demarcate the genome-wide boundaries of transcription units and might be a promising alternative to the hybrid approach applied in this work.

4.6.2. Differential expression testing by RNA-seq

The amount of differentially expressed genes matched well with the expectations, in that only 70 genes were significantly DE between Mal-MM and Mal-MM-TE, whereas 546 genes were found to be significantly DE between Mal-MM and Glc-CM, which differ in the composition of the cultivation media to a much greater extends than Mal-MM and Mal-MM-TE. It is remarkably that the provided differences in media conditions were so well reflected within the gene expression data, given the lack of replicates and the consequential conservative test statistic of the `DESeq` software. These findings approve the applicability of zero-replicate experiments on the one hand, and highlight the scalability of RNA-seq experiments on the other, in the sense that the discriminatory power of DE testing can be *bought* by adding further replicates to the experimental design.

4.6.3. Short assessment of computational methods for bacterial RNA-seq analysis

The major bioinformatic challenges involved in bacterial RNA-seq studies can generally be categorized into five tasks:

1. Transcriptome *de novo* assembly
2. Short read reference mapping
3. Expression quantification (normalization and summarization)
4. Differential expression testing
5. Annotation improvement

For the reason that RNA-seq is a rather new analytical method, it is consequential that no clear *gold standard* in terms of computational data analysis has been reached yet. This is reflected by a plurality of available software tools, e.g. for differential expression testing, which ranges from rather established tools, such as `DESeq` [Anders & Huber, 2010], `edgerR` [Robinson *et al.*, 2010], and `Cufflinks` [Trapnell *et al.*, 2010], up to latest releases, such as `NOISeq` [Tarazona *et al.*, 2011] and `GENE-Counter` [Cumbie *et al.*, 2011]. Likewise, new transcriptome assembly algorithms, short read mapping procedures, and expression quantification methods are constantly developed and improved, respectively. Advances in these categories have recently been extensively reviewed and indicate in-depth addressing by the bioinformatics community [Garber *et al.*, 2011, Chen *et al.*, 2011b].

In contrast, computational tools that aid in improving the genome annotation based on RNA-seq data are sorely lacking, which necessitates the development of custom scripts as described in **Section 2.6.6** or time consuming manual annotation [Wurtzel *et al.*, 2010]. The reason for the absence of appropriate tools is certainly based on the high degree of complexity that can rapidly arise from ambiguous RNA signals in diverse genomic contexts. Furthermore, the annotation problem can not be framed into mathematical terms right away, but rather needs to be broken down into smaller tasks e.g. as listed in **Section 3.4.2**. In addition, oftentimes no defined

optimal solution exists for a problem, e.g. the determination of a transcript's length which exhibits alternative TSSs. Moreover, a general RNA-seq annotation program has to cope with data derived from different cDNA libraries and, thus, needs to adjust for varying sequencing depths and read types (single or paired-end).

In the course of this study, several scripts were developed to solve the individual smaller tasks mentioned above. However, manual inspection and curation of the results was always necessary which might indicate that a semi-automated GUI-based application would hold the most benefit.

Another improvable aspect of current RNA-seq tools is the necessity to employ multiple programs of different origin in order to achieve the desired results. To the best of the author's knowledge, no integrated toolsuite or processing pipeline for complete RNA-seq analyses is currently available in the public domain. Nevertheless, a trend into this direction is observable and several developers already published more sophisticated analysis systems [Trapnell *et al.*, 2010, Howe *et al.*, 2011].

Chapter 5.
5 Conclusions and Outlook

This thesis was aimed at answering four major research questions related to *Actinoplanes* genomics and transcriptomics as stated in **Section 1.9**. The two foremost objectives were accomplished by establishing the complete 9.24 Mb genome sequence of the industrial acarbose producer *Actinoplanes* sp. SE50/110 and the annotation of more than 8,400 genomic features thereon. On that basis, special genes and gene clusters, such as the AICE, the potential antibiotic gene clusters and, of course, the acarbose biosynthetic gene cluster itself were analyzed with respect to their potential utilization in future strain enhancements. Lastly, successful transcriptome experiments were conducted using the novel RNA-sequencing technology. This method clearly demonstrated its added benefit over microarrays by not only permitting substantiated expression analysis, but also allowing for annotation improvements, such as the detection of non-coding RNAs and antisense transcripts. In conclusion, all planned objectives of this thesis were accomplished well beyond standard expectations.

The establishment of the complete genome sequence of the acarbose producer *Actinoplanes* sp. SE50/110 is an important achievement on the way towards rational optimization of the acarbose production through targeted genetic engineering. In this process, the identified AICE may serve as the basis for a future transformation system for this strain and other *Actinoplanes* spp. Furthermore, this work provides the first sequenced genome of the genus *Actinoplanes*, which will serve as the reference for future genome analysis and sequencing projects in this field.

With the complete genome sequence at hand, future transcriptome studies on *Actinoplanes* sp. SE50/110 should be conduction in order to systematically analyze the transcriptional effects of different carbon sources, which should be individually supplied to a reference cultivation media. In conjunction with measurements of CDW, acarbose, and possibly component C, these results would provide novel insights into the carbon metabolism of *Actinoplanes* sp. SE50/110 and might help to identify further potential target genes for later genetic manipulations with the aim of increasing acarbose yields. In order to cut costs, these experiments might also be conducted us-

ing cheaper microarrays, which should be designed to also include the novel features that were annotated on the basis of the performed RNA-seq experiments.

Moreover, recording transcriptional changes over the course of a cultivation might be of elevated interest in order to determine the exact timepoints at which acarbose production is initiated through an up-regulation of its biosynthetic gene cluster and when it is repressed again. These time series analysis could thus shed more light on the questionable relation between expression level of *acb* genes and the actual acarbose production rate, which might not be linear as indicated in this study.

Many possible knock-out and gene deletion targets were proposed in this study. In addition, several options for the construction of genetic modification systems were discussed, in particular the novel AICE pACPL. Taking these developments into account, the next logical step would be the development of a functional transformation system for *Actinoplanes* sp. SE50/110. With this at hand, the proposed trehalose synthase genes should be subject to first knock-out studies in order to solve the main problem in acarbose fermentation – the formation of component C. Future experiments could then be aimed at increasing the acarbose yields. For instance, knock-outs of the proposed acarbose importers might increase the extracellular concentration of acarbose, since it could not be re-imported into the cell.

By providing novel insights into the enzymatic equipment of *Actinoplanes* sp. SE50/110, previously unknown NRPS/PKS gene clusters were identified, potentially encoding new antibiotics and other bioactive compounds, which might be of pharmacologic interest. It might also be financially rewarding to examine the putative antibiotic gene clusters in more detail. In particular cACPL_1, putatively encoding a thiazomycin biosynthesis gene cluster, might be of foremost interest because of its pharmaceutical applicability. Moreover, no high yielding strain has been developed for this group of antibiotics yet, which is a clear advantage for *Actinoplanes* sp. SE50/110, since its genome sequence is known and genetic tools are underway.

It was also shown that the hybrid NRPS/PKS gene cluster cACPL_4 was highly induced under cultivation conditions with maltose as carbon source (Mal-MM and Mal-MM-TE) and mostly silent in the Glc-CM condition. This expressional pattern is highly similar to that observed for the acarbose gene cluster, which suggests a parallel expression of both clusters. Following up on this, future efforts in deleting or silencing cACPL_4 might result in an increased acarbose production, as it is likely that considerable amounts of cellular resources can then be utilized by the *acb*- instead of the cACPL_4-cluster.

Finally, the established full genome sequence of *Actinoplanes* sp. SE50/110 as well as the conducted transcriptomics experiments form the basis for further analytical exploration of the organism with other *omics*-technologies. In particular, recently initiated proteomics studies already unraveled important insights into the extra- and intracellular proteome of *Actinoplanes* sp. SE50/110, which would have been impossible without the knowledge of all coding sequences that were established in this study. Likewise, the presented genome sequence and its encoded genes are prerequisites for future metabolomic studies and *in silico* modeling in the sense of applied systems biology.

Bibliography

[Ahn, 2011] Ahn, S. (2011). Introduction to bioinformatics: sequencing technology. *Asia Pacific Allergy*, 1(2), 93–97. PMID: 22053303 PMCID: 3206250.

[Almeida *et al.*, 2004] Almeida, L. G. P., Paixão, R., Souza, R. C., da Costa, G. C., Barrientos, F. J. A., dos Santos, M. T., de Almeida, D. F., & Vasconcelos, A. T. R. (2004). A system for automated bacterial (genome) integrated Annotation-SABIA. *Bioinformatics*, 20(16), 2832 –2833.

[Altschul *et al.*, 1990] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), 403–410. PMID: 2231712.

[Anand *et al.*, 2010] Anand, S., Prasad, M. V. R., Yadav, G., Kumar, N., Shehara, J., Ansari, M. Z., & Mohanty, D. (2010). SBSPKS: structure based sequence analysis of polyketide synthases. *Nucleic Acids Research*, 38(Web Server), W487–W496.

[Anders & Huber, 2010] Anders, S. & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11(10), R106. PMID: 20979621.

[Andrews *et al.*, 2003] Andrews, S. C., Robinson, A. K., & Rodríguez-Quiñones, F. (2003). Bacterial iron homeostasis. *FEMS Microbiology Reviews*, 27(2-3), 215–237. PMID: 12829269.

[Apweiler *et al.*, 2004] Apweiler, R., Bairoch, A., & Wu, C. H. (2004). Protein sequence databases. *Current Opinion in Chemical Biology*, 8(1), 76–80.

[Aretz *et al.*, 2000] Aretz, W., Meiwes, J., Seibert, G., Vobis, G., & Wink, J. (2000). Friulimicins: novel lipopeptide antibiotics with peptidoglycan synthesis inhibiting activity from *Actinoplanes friuliensis* sp. nov. I. taxonomic studies of the producing microorganism and fermentation. *The Journal of Antibiotics*, 53(8), 807–815. PMID: 11079803.

[Avonce *et al.*, 2006] Avonce, N., Mendoza-Vargas, A., Morett, E., & Iturriaga, G. (2006). Insights on the evolution of trehalose biosynthesis. *BMC Evolutionary Biology*, 6, 109. PMID: 17178000.

[Badger & Olsen, 1999] Badger, J. H. & Olsen, G. J. (1999). CRITICA: coding region identification tool invoking comparative analysis. *Molecular Biology and Evolution*, 16(4), 512–524. PMID: 10331277.

[Bainbridge *et al.*, 2006] Bainbridge, M. N., Warren, R. L., Hirst, M., Romanuik, T., Zeng, T., Go, A., Delaney, A., Griffith, M., Hickenbotham, M., Magrini, V., Mardis, E. R., Sadar, M. D., Siddiqui, A. S., Marra, M. A., & Jones, S. J. M. (2006). Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach. *BMC Genomics*, 7, 246. PMID: 17010196.

[Bairoch, 2000] Bairoch, A. (2000). The ENZYME database in 2000. *Nucleic Acids Research*, 28(1), 304 –305.

[Baltz, 1998] Baltz, R. H. (1998). New genetic methods to improve secondary metabolite production in *Streptomyces*. *Journal of Industrial Microbiology and Biotechnology*, 20, 360–363.

[Baltz, 2001] Baltz, R. H. (2001). Genetic methods and strategies for secondary metabolite yield improvement in actinomycetes. *Antonie Van Leeuwenhoek*, 79(3-4), 251–259. PMID: 11816967.

[Baltz, 2011] Baltz, R. H. (2011). Strain improvement in actinomycetes in the postgenomic era. *Journal of Industrial Microbiology & Biotechnology*, 38(6), 657–666. PMID: 21253811.

[Baltz & Hosted, 1996] Baltz, R. H. & Hosted, T. J. (1996). Molecular genetic methods for improving secondary-metabolite production in actinomycetes. *Trends in Biotechnology*, 14(7), 245–250. PMID: 8771797.

[Banerjee *et al.*, 2006] Banerjee, S., Chalissery, J., Bandey, I., & Sen, R. (2006). Rho-dependent transcription termination: More questions than answers. *Journal of microbiology (Seoul, Korea)*, 44(1), 11–22. PMID: 16554712 PMCID: 1838574.

[Bashir *et al.*, 2008] Bashir, A., Volik, S., Collins, C., Bafna, V., & Raphael, B. J. (2008). Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. *PLoS Computational Biology*, 4(4), e1000051. PMID: 18404202.

[Bayer AG, 2011] Bayer AG (2011). *Geschäftsbericht 2010*. Geschäftsbericht, Bayer AG, Leverkusen, Germany.

[Bearden *et al.*, 1998] Bearden, S. W., Staggs, T. M., & Perry, R. D. (1998). An ABC transporter system of *Yersinia pestis* allows utilization of chelated iron by *Escherichia coli* SAB11. *Journal of Bacteriology*, 180(5), 1135–1147. PMID: 9495751.

[Bendtsen *et al.*, 2005] Bendtsen, J. D., Kiemer, L., Fausbøll, A., & Brunak, S. (2005). Non-classical protein secretion in bacteria. *BMC Microbiology*, 5, 58. PMID: 16212653.

[Bendtsen *et al.*, 2004] Bendtsen, J. D., Nielsen, H., von Heijne, G., & Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. *Journal of Molecular Biology*, 340(4), 783–795. PMID: 15223320.

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

[Besemer & Borodovsky, 1999] Besemer, J. & Borodovsky, M. (1999). Heuristic approach to deriving models for gene finding. *Nucleic Acids Research*, 27(19), 3911–3920. PMID: 10481031.

[Besemer *et al.*, 2001] Besemer, J., Lomsadze, A., & Borodovsky, M. (2001). GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. implications for finding sequence motifs in regulatory regions. *Nucleic Acids Research*, 29(12), 2607–2618. PMID: 11410670.

[Bibb *et al.*, 1984] Bibb, M. J., Findlay, P. R., & Johnson, M. W. (1984). The relationship between base composition and codon usage in bacterial genes and its use for the simple and reliable identification of protein-coding sequences. *Gene*, 30(1-3), 157–166. PMID: 6096212.

[Birol *et al.*, 2009] Birol, I., Jackman, S. D., Nielsen, C. B., Qian, J. Q., Varhol, R., Stazyk, G., Morin, R. D., Zhao, Y., Hirst, M., Schein, J. E., Horsman, D. E., Connors, J. M., Gascoyne, R. D., Marra, M. A., & Jones, S. J. M. (2009). *De novo* transcriptome assembly with ABySS. *Bioinformatics (Oxford, England)*, 25(21), 2872–2877. PMID: 19528083.

[Blom *et al.*, 2009] Blom, J., Albaum, S. P., Doppmeier, D., Pühler, A., Vorhölter, F., Zakrzewski, M., & Goesmann, A. (2009). EDGAR: a software framework for the comparative analysis of prokaryotic genomes. *BMC Bioinformatics*, 10, 154. PMID: 19457249.

[Blom *et al.*, 2011] Blom, J., Jakobi, T., Doppmeier, D., Jaenicke, S., Kalinowski, J., Stoye, J., & Goesmann, A. (2011). Exact and complete short-read alignment to microbial genomes using graphics processing unit programming. *Bioinformatics (Oxford, England)*, 27(10), 1351–1358. PMID: 21450712.

[Boakes *et al.*, 2010] Boakes, S., Appleyard, A. N., Cortés, J., & Dawson, M. J. (2010). Organization of the biosynthetic genes encoding deoxyactagardine B (DAB), a new lantibiotic produced by *Actinoplanes liguriae* NCIMB41362. *The Journal of Antibiotics*, 63(7), 351–358. PMID: 20520597.

[Boakes *et al.*, 2009] Boakes, S., Cortés, J., Appleyard, A. N., Rudd, B. A. M., & Dawson, M. J. (2009). Organization of the genes encoding the biosynthesis of

actagardine and engineering of a variant generation system. *Molecular Microbiology*, 72(5), 1126–1136. PMID: 19400806.

[Borovok *et al.*, 2006] Borovok, I., Gorovitz, B., Schreiber, R., Aharonowitz, Y., & Cohen, G. (2006). Coenzyme B12 controls transcription of the *Streptomyces* class Ia ribonucleotide reductase *nrdABS* operon via a riboswitch mechanism. *Journal of Bacteriology*, 188(7), 2512 –2520.

[Borovok *et al.*, 2004] Borovok, I., Gorovitz, B., Yanku, M., Schreiber, R., Gust, B., Chater, K., Aharonowitz, Y., & Cohen, G. (2004). Alternative oxygen-dependent and oxygen-independent ribonucleotide reductases in *Streptomyces*: cross-regulation and physiological role in response to oxygen limitation. *Molecular Microbiology*, 54(4), 1022–1035. PMID: 15522084.

[Bottino & Trucco, 2005] Bottino, R. & Trucco, M. (2005). Multifaceted therapeutic approaches for a multigenic disease. *Diabetes*, 54 Suppl 2, S79–86. PMID: 16306345.

[Brekasis & Paget, 2003] Brekasis, D. & Paget, M. S. B. (2003). A novel sensor of NADH/NAD+ redox poise in *Streptomyces coelicolor* A3(2). *The EMBO Journal*, 22(18), 4856–4865. PMID: 12970197.

[Brunkhorst & Schneider, 2005] Brunkhorst, C. & Schneider, E. (2005). Characterization of maltose and maltotriose transport in the acarbose-producing bacterium *Actinoplanes* sp. *Research in Microbiology*, 156(8), 851–857. PMID: 15939574.

[Brunkhorst *et al.*, 2005] Brunkhorst, C., Wehmeier, U. F., Piepersberg, W., & Schneider, E. (2005). The *acbH* gene of *Actinoplanes* sp. encodes a solute receptor with binding activities for acarbose and longer homologs. *Research in Microbiology*, 156(3), 322–327. PMID: 15808935.

[Bullard *et al.*, 2010] Bullard, J. H., Purdom, E., Hansen, K. D., & Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, 11, 94. PMID: 20167110.

[Burrus & Waldor, 2004] Burrus, V. & Waldor, M. K. (2004). Shaping bacterial genomes with integrative and conjugative elements. *Research in Microbiology*, 155(5), 376–386. PMID: 15207870.

[Cantarel *et al.*, 2008] Cantarel, B. L., Korf, I., Robb, S. M., Parra, G., Ross, E., Moore, B., Holt, C., Sánchez Alvarado, A., & Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, 18(1), 188 –196.

[Cao *et al.*, 2007] Cao, J., Woodhall, M. R., Alvarez, J., Cartron, M. L., & Andrews, S. C. (2007). EfeUOB (YcdNOB) is a tripartite, acid-induced and CpxAR-regulated, low-pH Fe^{2+} transporter that is cryptic in *Escherichia coli* K-12 but functional in *E. coli* O157:H7. *Molecular Microbiology*, 65(4), 857–875. PMID: 17627767.

[Carlson *et al.*, 2006] Carlson, J. M., Chakravarty, A., Khetani, R. S., & Gross, R. H. (2006). Bounded search for *de novo* identification of degenerate cis-regulatory elements. *BMC Bioinformatics*, 7, 254. PMID: 16700920.

[Carrondo, 2003] Carrondo, M. A. (2003). Ferritins, iron uptake and storage from the bacterioferritin viewpoint. *The EMBO Journal*, 22(9), 1959–1968. PMID: 12727864.

[Caspary & Graf, 1979] Caspary, W. F. & Graf, S. (1979). Inhibition of human intestinal alpha-glucosidehydrolases by a new complex oligosaccharide. *Research in Experimental Medicine. Zeitschrift Für Die Gesamte Experimentelle Medizin Einschliesslich Experimenteller Chirurgie*, 175(1), 1–6. PMID: 441522.

[Caspary & Kalisch, 1979] Caspary, W. F. & Kalisch, H. (1979). Effect of alpha-glucosidehydrolase inhibition and intestinal absorption of sucrose, water, and sodium in man. *Gut*, 20(9), 750–755. PMID: 387540.

[Cavalleri *et al.*, 1984] Cavalleri, B., Pagani, H., Volpe, G., Selva, E., & Parenti, F. (1984). A-16686, a new antibiotic from *Actinoplanes*. I. fermentation, isolation and preliminary physico-chemical characteristics. *The Journal of Antibiotics*, 37(4), 309–317. PMID: 6547132.

[Cayuela *et al.*, 2003] Cayuela, M. L., Elías-Arnanz, M., Peñalver-Mellado, M., Padmanabhan, S., & Murillo, F. J. (2003). The *Stigmatella aurantiaca* homolog of *Myxococcus xanthus* high-mobility-group A-type transcription factor CarD: insights into the functional modules of CarD and their distribution in bacteria. *Journal of Bacteriology*, 185(12), 3527–3537. PMID: 12775690.

[Chain *et al.*, 2009] Chain, P. S. G., Grafham, D. V., Fulton, R. S., FitzGerald, M. G., Hostetler, J., Muzny, D., Ali, J., Birren, B., Bruce, D. C., Buhay, C., Cole, J. R., Ding, Y., Dugan, S., Field, D., Garrity, G. M., Gibbs, R., Graves, T., Han, C. S., Harrison, S. H., Highlander, S., Hugenholtz, P., Khouri, H. M., Kodira, C. D., Kolker, E., Kyrpides, N. C., Lang, D., Lapidus, A., Malfatti, S. A., Markowitz, V., Metha, T., Nelson, K. E., Parkhill, J., Pitluck, S., Qin, X., Read, T. D., Schmutz, J., Sozhamannan, S., Sterk, P., Strausberg, R. L., Sutton, G., Thomson, N. R., Tiedje, J. M., Weinstock, G., Wollam, A., Consortium, G. S. C. H. M. P. J., & Detter, J. C. (2009). Genome project standards in a new era of sequencing. *Science*, 326(5950), 236–237.

[Challis *et al.*, 2000] Challis, G. L., Ravel, J., & Townsend, C. A. (2000). Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chemistry & Biology*, 7(3), 211–224.

[Chen *et al.*, 2011a] Chen, G., Li, R., Shi, L., Qi, J., Hu, P., Luo, J., Liu, M., & Shi, T. (2011a). Revealing the missing expressed genes beyond the human reference genome by RNA-Seq. *BMC Genomics*, 12(1), 590. PMID: 22133125.

[Chen *et al.*, 2011b] Chen, G., Wang, C., & Shi, T. (2011b). Overview of available methods for diverse RNA-Seq data analyses. *Science China. Life Sciences*, 54(12), 1121–1128. PMID: 22227904.

[Chinault & Carbon, 1979] Chinault, A. C. & Carbon, J. (1979). Overlap hybridization screening: isolation and characterization of overlapping DNA fragments surrounding the *leu2* gene on yeast chromosome III. *Gene*, 5(2), 111–126. PMID: 376402.

[Claudel-Renard *et al.*, 2003] Claudel-Renard, C., Chevalet, C., Faraut, T., & Kahn, D. (2003). Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Research*, 31(22), 6633 –6639.

[Cloonan *et al.*, 2008] Cloonan, N., Forrest, A. R. R., Kolle, G., Gardiner, B. B. A., Faulkner, G. J., Brown, M. K., Taylor, D. F., Steptoe, A. L., Wani, S., Bethel, G., Robertson, A. J., Perkins, A. C., Bruce, S. J., Lee, C. C., Ranade, S. S., Peckham, H. E., Manning, J. M., McKernan, K. J., & Grimmond, S. M. (2008). Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods*, 5(7), 613–619. PMID: 18516046.

[Cole *et al.*, 2009] Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., Kulam-Syed-Mohideen, A. S., McGarrell, D. M., Marsh, T., Garrity, G. M., & Tiedje, J. M. (2009). The ribosomal database project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Research*, 37(Database issue), D141–145. PMID: 19004872.

[Cornelis *et al.*, 2011] Cornelis, P., Wei, Q., Andrews, S. C., & Vinckx, T. (2011). Iron homeostasis and management of oxidative stress response in bacteria. *Metallomics: Integrated Biometal Science*, 3(6), 540–549. PMID: 21566833.

[Costa *et al.*, 2010] Costa, V., Angelini, C., De Feis, I., & Ciccodicola, A. (2010). Uncovering the complexity of transcriptomes with RNA-Seq. *Journal of Biomedicine and Biotechnology*, 2010. PMID: 20625424 PMCID: 2896904.

[Couch, 1950] Couch, J. N. (1950). *Actinoplanes*, a new genus of the actinomycetales. *Journal of the Elisha Mitchell Scientific Society*, 66.

[Couch, 1963] Couch, J. N. (1963). Some new genera and species of the actinoplanaceae. 79, 53–70.

[Coulson, 1994] Coulson, A. (1994). High-performance searching of biosequence databases. *Trends in Biotechnology*, 12(3), 76–80. PMID: 7764827.

[Crooks *et al.*, 2004] Crooks, G. E., Hon, G., Chandonia, J., & Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Research*, 14(6), 1188–1190. PMID: 15173120 PMCID: 419797.

[Croucher & Thomson, 2010] Croucher, N. J. & Thomson, N. R. (2010). Studying bacterial transcriptomes using RNA-seq. *Current Opinion in Microbiology*, 13(5), 619–624. PMID: 20888288 PMCID: 3025319.

[Crueger, 2000] Crueger, A. (2000). Immobilized cells of *Actinoplanes acarbosefaciens* SE50/110, method for immobilizing same and use of the immobilisate for the production of acarbose. Patent. WO/2000/018901.

[Crueger *et al.*, 1998a] Crueger, A., Apeler, H., Schröder, W., Pape, H., Goeke, K., Piepersberg, W., Distler, J., Diaz-Guardamino Uribe, P. M., Jarling, M., & Stratmann, A. (1998a). Acarbose (Acb) cluster from *Actinoplanes* sp. SE 50/110.

[Crueger *et al.*, 1998b] Crueger, A., Piepersberg, W., Distler, J., & Stratmann, A. (1998b). Acarbose biosynthesis genes from *Actinoplanes* sp., process for the isolation thereof and the use thereof. Patent. US 5753501.

[Cumbie *et al.*, 2011] Cumbie, J. S., Kimbrel, J. A., Di, Y., Schafer, D. W., Wilhelm, L. J., Fox, S. E., Sullivan, C. M., Curzon, A. D., Carrington, J. C., Mockler, T. C., & Chang, J. H. (2011). GENE-counter: a computational pipeline for the analysis of RNA-Seq data for gene expression differences. *PloS One*, 6(10), e25279. PMID: 21998647.

[Das *et al.*, 2005] Das, A., Silaghi-Dumitrescu, R., Ljungdahl, L. G., & Kurtz, D. M. (2005). Cytochrome *bd* oxidase, oxidative stress, and dioxygen tolerance of the strictly anaerobic bacterium *Moorella thermoacetica*. *J. Bacteriol.*, 187(6), 2020–2029.

[Dauter *et al.*, 1999] Dauter, Z., Dauter, M., Brzozowski, A. M., Christensen, S., Borchert, T. V., Beier, L., Wilson, K. S., & Davies, G. J. (1999). X-ray structure of novamyl, the five-domain "maltogenic" alpha-amylase from *Bacillus stearothermophilus*: maltose and acarbose complexes at 1.7A resolution. *Biochemistry*, 38(26), 8385–8392. PMID: 10387084.

[Debut *et al.*, 2006] Debut, A. J., Dumay, Q. C., Barabote, R. D., & Saier, Milton H, J. (2006). The iron/lead transporter superfamily of Fe/Pb2+ uptake systems. *Journal of Molecular Microbiology and Biotechnology*, 11(1-2), 1–9. PMID: 16825785.

[Delcher *et al.*, 1999] Delcher, A. L., Harmon, D., Kasif, S., White, O., & Salzberg, S. L. (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Research*, 27(23), 4636–4641. PMID: 10556321.

[Deslandes *et al.*, 2007] Deslandes, V., Nash, J. H. E., Harel, J., Coulton, J. W., & Jacques, M. (2007). Transcriptional profiling of *Actinobacillus pleuropneumoniae* under iron-restricted conditions. *BMC Genomics*, 8, 72. PMID: 17355629.

[Deutscher, 2003] Deutscher, M. P. (2003). Degradation of stable RNA in bacteria. *Journal of Biological Chemistry*, 278(46), 45041 –45044.

[Deutscher, 2006] Deutscher, M. P. (2006). Degradation of RNA in bacteria: comparison of mRNA and stable RNA. *Nucleic Acids Research*, 34(2), 659 –666.

[Dohm *et al.*, 2008] Dohm, J. C., Lottaz, C., Borodina, T., & Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, 36(16), e105. PMID: 18660515.

[Drake *et al.*, 2007] Drake, E. J., Cao, J., Qu, J., Shah, M. B., Straubinger, R. M., & Gulick, A. M. (2007). The 1.8 A crystal structure of PA2412, an MbtH-like protein from the pyoverdine cluster of *Pseudomonas aeruginosa*. *The Journal of Biological Chemistry*, 282(28), 20425–20434. PMID: 17502378.

[Dreier & Khosla, 2000] Dreier, J. & Khosla, C. (2000). Mechanistic analysis of a type II polyketide synthase. role of conserved residues in the β-Ketoacyl Synthase-Chain length factor heterodimer. *Biochemistry*, 39(8), 2088–2095.

[Drepper & Pape, 1996] Drepper, A. & Pape, H. (1996). Acarbose 7-phosphotransferase from *Actinoplanes* sp.: purification, properties, and possible physiological function. *The Journal of Antibiotics*, 49(7), 664–668. PMID: 8784428.

[Du & Lou, 2010] Du, L. & Lou, L. (2010). PKS and NRPS release mechanisms. *Natural Product Reports*, 27(2), 255–278. PMID: 20111804.

[Du *et al.*, 2001] Du, L., Sánchez, C., & Shen, B. (2001). Hybrid peptide-polyketide natural products: biosynthesis and prospects toward engineering novel molecules. *Metabolic Engineering*, 3(1), 78–95. PMID: 11162234.

[Edgar, 2004a] Edgar, R. C. (2004a). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5, 113. PMID: 15318951.

[Edgar, 2004b] Edgar, R. C. (2004b). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5), 1792–1797. PMID: 15034147.

[Eichler *et al.*, 2004] Eichler, E. E., Clark, R. A., & She, X. (2004). An assessment of the sequence gaps: Unfinished business in a finished human genome. *Nat Rev Genet*, 5(5), 345–354.

[Elías-Arnanz *et al.*, 2010] Elías-Arnanz, M., Padmanabhan, S., & Murillo, F. J. (2010). The regulatory action of the myxobacterial CarD/CarG complex: a bacterial enhanceosome? *FEMS Microbiology Reviews*, 34(5), 764–778. PMID: 20561058.

[Escolar *et al.*, 1999] Escolar, L., Perez-Martin, J., & de Lorenzo, V. (1999). Opening the iron box: Transcriptional metalloregulation by the fur protein. *J. Bacteriol.*, 181(20), 6223–6229.

[Esteve-Codina *et al.*, 2011] Esteve-Codina, A., Kofler, R., Palmieri, N., Bussotti, G., Notredame, C., & Pérez-Enciso, M. (2011). Exploring the gonad transcriptome of two extreme male pigs with RNA-seq. *BMC Genomics*, 12, 552. PMID: 22067327.

[Ewing & Green, 1998] Ewing, B. & Green, P. (1998). Base-calling of automated sequencer traces using phred. II. error probabilities. *Genome Research*, 8(3), 186–194. PMID: 9521922.

[Ewing *et al.*, 1998] Ewing, B., Hillier, L., Wendl, M. C., & Green, P. (1998). Base-calling of automated sequencer traces using phred. I. accuracy assessment. *Genome Research*, 8(3), 175–185. PMID: 9521921.

[Farina & Bradley, 1970] Farina, G. & Bradley, S. G. (1970). Reassociation of deoxyribonucleic acids from *Actinoplanes* and other actinomycetes. *Journal of Bacteriology*, 102(1), 30–35. PMID: 5437730 PMCID: 284966.

[Felnagle *et al.*, 2008] Felnagle, E. A., Jackson, E. E., Chan, Y. A., Podevels, A. M., Berti, A. D., McMahon, M. D., & Thomas, M. G. (2008). Nonribosomal peptide synthetases involved in the production of medically relevant natural products. *Molecular Pharmaceutics*, 5(2), 191–211.

[Felsenstein, 1985] Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. 3, 783–791.

[Finn *et al.*, 2009] Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L. L., Eddy, S. R., & Bateman, A. (2009). The Pfam protein families database. *Nucleic Acids Research*, 38(Database), D211–D222.

[Fraser *et al.*, 2002] Fraser, C. M., Eisen, J. A., Nelson, K. E., Paulsen, I. T., & Salzberg, S. L. (2002). The value of complete microbial genome sequencing (You get what you pay for). *J. Bacteriol.*, 184(23), 6403–6405.

[Frey *et al.*, 2008] Frey, U. H., Bachmann, H. S., Peters, J., & Siffert, W. (2008). PCR-amplification of GC-rich regions: 'slowdown PCR'. *Nat. Protocols*, 3(8), 1312–1317.

[Frishman *et al.*, 1998] Frishman, D., Mironov, A., Mewes, H. W., & Gelfand, M. (1998). Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Research*, 26(12), 2941–2947. PMID: 9611239.

[Frommer *et al.*, 1977a] Frommer, W., Junge, B., Keup, U., Mueller, L., & Schmidt, D. (1977a). Amino sugar derivatives. Patent. DE 2347782 (US patent 4,062,950).

[Frommer *et al.*, 1979] Frommer, W., Junge, B., Müller, L., Schmidt, D., & Truscheit, E. (1979). Neue Enzyminhibitoren aus Mikroorganismen. *Planta Medica*, 35(3), 195–217. PMID: 432298.

[Frommer *et al.*, 1975] Frommer, W., Puls, W., Schäfer, D., & Schmidt, D. (1975). Glycoside-hydrolase enzyme inhibitors. Patent. DE 2064092 (US patent 3,876,766).

[Frommer *et al.*, 1977b] Frommer, W., Puls, W., & Schmidt, D. (1977b). Process for the production of a saccharase inhibitor. Patent. DE 2209834 (US patent 4,019,960).

[Gaasterland & Sensen, 1996] Gaasterland, T. & Sensen, C. W. (1996). Fully automated genome analysis that reflects user needs and preferences. a detailed introduction to the MAGPIE system architecture. *Biochimie*, 78(5), 302–310. PMID: 8905148.

[Gao & Zhang, 2008] Gao, F. & Zhang, C. (2008). Ori-Finder: a web-based system for finding *oriCs* in unannotated bacterial genomes. *BMC Bioinformatics*, 9, 79. PMID: 18237442.

[Garber *et al.*, 2011] Garber, M., Grabherr, M. G., Guttman, M., & Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods*, 8(6), 469–477. PMID: 21623353.

[García-Castellanos *et al.*, 2004] García-Castellanos, R., Mallorquí-Fernández, G., Marrero, A., Potempa, J., Coll, M., & Gomis-Rüth, F. X. (2004). On the transcriptional regulation of methicillin resistance. *Journal of Biological Chemistry*, 279(17), 17888 –17896.

[Gardner *et al.*, 2009] Gardner, P. P., Daub, J., Tate, J. G., Nawrocki, E. P., Kolbe, D. L., Lindgreen, S., Wilkinson, A. C., Finn, R. D., Griffiths-Jones, S., Eddy, S. R., & Bateman, A. (2009). Rfam: updates to the RNA families database. *Nucleic Acids Research*, 37(Database), D136–D140.

[Garg *et al.*, 1996] Garg, R. P., Vargo, C. J., Cui, X., & Kurtz, D M, J. (1996). A [2Fe-2S] protein encoded by an open reading frame upstream of the *Escherichia coli* bacterioferritin gene. *Biochemistry*, 35(20), 6297–6301. PMID: 8639572.

[Güell *et al.*, 2009] Güell, M., van Noort, V., Yus, E., Chen, W., Leigh-Bell, J., Michalodimitrakis, K., Yamada, T., Arumugam, M., Doerks, T., Kühner, S., Rode, M., Suyama, M., Schmidt, S., Gavin, A., Bork, P., & Serrano, L. (2009). Transcriptome complexity in a genome-reduced bacterium. *Science (New York, N.Y.)*, 326(5957), 1268–1271. PMID: 19965477.

[Ghaemmaghami *et al.*, 2003] Ghaemmaghami, S., Huh, W., Bower, K., Howson, R. W., Belle, A., Dephoure, N., O'Shea, E. K., & Weissman, J. S. (2003). Global analysis of protein expression in yeast. *Nature*, 425(6959), 737–741.

[Goeke *et al.*, 1996] Goeke, K., Drepper, A., & Pape, H. (1996). Formation of acarbose phosphate by a cell-free extract from the acarbose producer *Actinoplanes* sp. *The Journal of Antibiotics*, 49(7), 661–663. PMID: 8784426.

[Gordon *et al.*, 1998] Gordon, D., Abajian, C., & Green, P. (1998). Consed: a graphical tool for sequence finishing. *Genome Research*, 8(3), 195 –202.

[Gordon *et al.*, 2001] Gordon, D., Desmarais, C., & Green, P. (2001). Automated finishing with autofinish. *Genome Research*, 11(4), 614 –625.

[Gottesman, 2005] Gottesman, S. (2005). Micros for microbes: non-coding regulatory RNAs in bacteria. *Trends in Genetics: TIG*, 21(7), 399–404. PMID: 15913835.

[Griffiths-Jones, 2004] Griffiths-Jones, S. (2004). Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research*, 33(Database issue), D121–D124.

[Grohmann *et al.*, 2003] Grohmann, E., Muth, G., & Espinosa, M. (2003). Conjugative plasmid transfer in Gram-Positive bacteria. *Microbiol. Mol. Biol. Rev.*, 67(2), 277–301.

[Groves *et al.*, 2010] Groves, M. R., Ortiz, D., & Lucana, D. O. (2010). Adaptation to oxidative stress by Gram-positive bacteria: the redox sensing system HbpS-SenS-SenR from *Streptomyces reticuli*. *Applied Microbiology*, (pp. 33–42).

[Grunden *et al.*, 1999] Grunden, A. M., Self, W. T., Villain, M., Blalock, J. E., & Shanmugam, K. T. (1999). An analysis of the binding of repressor protein ModE to *modABCD* (molybdate transport) Operator/Promoter DNA of *Escherichia coli*. *Journal of Biological Chemistry*, 274(34), 24308 –24315.

[Gursinsky *et al.*, 2000] Gursinsky, T., Jäger, J., Andreesen, J. R., & Söhling, B. (2000). A *selDABC* cluster for selenocysteine incorporation in *Eubacterium acidaminophilum*. *Archives of Microbiology*, 174(3), 200–212.

[Ha *et al.*, 2008] Ha, H., Hwang, Y., & Choi, S. (2008). Application of conjugation using phiC31 att/int system for *Actinoplanes teichomyceticus*, a producer of teicoplanin. *Biotechnology Letters*, 30(7), 1233–1238. PMID: 18317703.

[Habermehl *et al.*, 2008] Habermehl, G., Hammann, P., Krebs, H., & Ternes, W. (2008). *Naturstoffchemie: Eine Einführung*. Springer-Lehrbuch. Springer.

[Hahn & Stachelhaus, 2004] Hahn, M. & Stachelhaus, T. (2004). Selective interaction between nonribosomal peptide synthetases is facilitated by short communication-mediating domains. *Proceedings of the National Academy of Sciences of the United States of America*, 101(44), 15585 –15590.

[Haiser *et al.*, 2008] Haiser, H. J., Karginov, F. V., Hannon, G. J., & Elliot, M. A. (2008). Developmentally regulated cleavage of tRNAs in the bacterium *Streptomyces coelicolor*. *Nucleic Acids Research*, 36(3), 732–741. PMID: 18084030.

[Hanozet *et al.*, 1981] Hanozet, G., Pircher, H. P., Vanni, P., Oesch, B., & Semenza, G. (1981). An example of enzyme hysteresis. the slow and tight interaction of some

fully competitive inhibitors with small intestinal sucrase. *The Journal of Biological Chemistry*, 256(8), 3703–3711. PMID: 6452453.

[Hardcastle & Kelly, 2010] Hardcastle, T. J. & Kelly, K. A. (2010). baySeq: empirical bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, 11, 422. PMID: 20698981.

[Hartmann *et al.*, 2009] Hartmann, R. K., Gössringer, M., Späth, B., Fischer, S., & Marchfelder, A. (2009). The making of tRNAs and more - RNase P and tRNase Z. *Progress in Molecular Biology and Translational Science*, 85, 319–368. PMID: 19215776.

[Hashimoto *et al.*, 2009] Hashimoto, S.-i., Qu, W., Ahsan, B., Ogoshi, K., Sasaki, A., Nakatani, Y., Lee, Y., Ogawa, M., Ametani, A., Suzuki, Y., Sugano, S., Lee, C. C., Nutter, R. C., Morishita, S., & Matsushima, K. (2009). High-resolution analysis of the 5'-end transcriptome using a next generation DNA sequencer. *PloS One*, 4(1), e4108. PMID: 19119315.

[He *et al.*, 2008] He, S., Liu, C., Skogerbø, G., Zhao, H., Wang, J., Liu, T., Bai, B., Zhao, Y., & Chen, R. (2008). NONCODE v2.0: decoding the non-coding. *Nucleic Acids Research*, 36(Database issue), D170–D172. PMID: 18000000 PMCID: 2238973.

[Heiker *et al.*, 1981] Heiker, F. R., Böshagen, H., Junge, B., Müller, L., & Stoltefuß, J. (1981). Studies designed to localize the essential structural unit of glycosidehydrolase inhibitors of the acarbose type. In C. E (Ed.), *first international symposium on acarbose* (pp. 137–141). Amsterdam: Excerpta Medica.

[Hemker *et al.*, 2001] Hemker, M., Stratmann, A., Goeke, K., Schröder, W., Lenz, J., Piepersberg, W., & Pape, H. (2001). Identification, cloning, expression, and characterization of the extracellular acarbose-modifying glycosyltransferase, AcbD, from *Actinoplanes* sp. strain SE50. *Journal of Bacteriology*, 183(15), 4484–4492. PMID: 11443082 PMCID: 95342.

[Hendrickson & Lawrence, 2007] Hendrickson, H. & Lawrence, J. G. (2007). Mutational bias suggests that replication termination occurs near the dif site, not at Ter sites. *Molecular Microbiology*, 64(1), 42–56. PMID: 17376071.

[Henke *et al.*, 1997] Henke, W., Herdel, K., Jung, K., Schnorr, D., & Loening, S. A. (1997). Betaine improves the PCR amplification of GC-rich DNA sequences. *Nucleic Acids Research*, 25(19), 3957–3958. PMID: 9380524.

[Hert *et al.*, 2008] Hert, D. G., Fredlake, C. P., & Barron, A. E. (2008). Advantages and limitations of next-generation sequencing technologies: A comparison of electrophoresis and non-electrophoresis methods. *ELECTROPHORESIS*, 29(23), 4618–4626.

[Höfs *et al.*, 2000] Höfs, R., Walker, M., & Zeeck, A. (2000). Hexacyclinic acid, a polyketide from *Streptomyces* with a novel carbon skeleton. *Angewandte Chemie International Edition*, 39(18), 3258–3261.

[Hosted *et al.*, 2005] Hosted, Thomas J, J., Wang, T., & Horan, A. C. (2005). Characterization of the *Micromonospora rosaria* pMR2 plasmid and development of a high G+C codon optimized integrase for site-specific integration. *Plasmid*, 54(3), 249–258. PMID: 16024079.

[Howe *et al.*, 2011] Howe, E. A., Sinha, R., Schlauch, D., & Quackenbush, J. (2011). RNA-Seq analysis in MeV. *Bioinformatics (Oxford, England)*, 27(22), 3209–3210. PMID: 21976420.

[Hyatt *et al.*, 2010] Hyatt, D., Chen, G., LoCascio, P., Land, M., Larimer, F., & Hauser, L. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, 11(1), 119.

[IDF, 2009] IDF (2009). *Diabetes Atlas, 4th Ed.* Brussels, Belgium: International Diabetes Federation.

[Ishikawa & Hotta, 1999] Ishikawa, J. & Hotta, K. (1999). FramePlot: a new implementation of the frame analysis for predicting protein-coding regions in bacterial DNA with a high G + C content. *FEMS Microbiology Letters*, 174(2), 251–253. PMID: 10339816.

[Jarling *et al.*, 2004a] Jarling, M., Bartkowiak, K., Pape, H., & Meinhardt, F. (2004a). The genome of phiAsp2, an *Actinoplanes* infecting phage. *Virus Genes*, 29(1), 117–129. PMID: 15215690.

[Jarling *et al.*, 2004b] Jarling, M., Bartkowiak, K., Robenek, H., Pape, H., & Meinhardt, F. (2004b). Isolation of phages infecting *Actinoplanes* SN223 and characterization of two of these viruses. *Applied Microbiology and Biotechnology*, 64(2), 250–254. PMID: 14586581.

[Jayasuriya *et al.*, 2007] Jayasuriya, H., Herath, K., Ondeyka, J. G., Zhang, C., Zink, D. L., Brower, M., Gailliot, F. P., Greene, J., Birdsall, G., Venugopal, J., Ushio, M., Burgess, B., Russotti, G., Walker, A., Hesse, M., Seeley, A., Junker, B., Connors, N., Salazar, O., Genilloud, O., Liu, K., Masurekar, P., Barrett, J. F., & Singh, S. B. (2007). Isolation and structure elucidation of thiazomycin. *J Antibiot*, 60(9), 554–564.

[Jukes & Cantor, 1969] Jukes, T. & Cantor, C. (1969). Evolution of protein molecules. In *Mammalian Protein Metabolism* (pp. pp. 21–132). New York: Academic Press.

[Jung *et al.*, 2009] Jung, H., Jeya, M., Kim, S., Moon, H., Kumar Singh, R., Zhang, Y., & Lee, J. (2009). Biosynthesis, biotechnological production, and application of

teicoplanin: current state and perspectives. *Applied Microbiology and Biotechnology*, 84(3), 417–428. PMID: 19609520.

[Jung *et al.*, 2008] Jung, H., Kim, S., Prabhu, P., Moon, H., Kim, I., & Lee, J. (2008). Optimization of culture conditions and scale-up to plant scales for teicoplanin production by *Actinoplanes teichomyceticus*. *Applied Microbiology and Biotechnology*, 80(1), 21–27. PMID: 18542948.

[Kanehisa & Goto, 2000] Kanehisa, M. & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1), 27–30. PMID: 10592173.

[Kanehisa *et al.*, 2006] Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K. F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., & Hirakawa, M. (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research*, 34(Database issue), D354–357. PMID: 16381885.

[Kataoka *et al.*, 1994] Kataoka, M., Kiyose, Y. M., Michisuji, Y., Horiguchi, T., Seki, T., & Yoshida, T. (1994). Complete nucleotide sequence of the *Streptomyces nigrifaciens* plasmid, pSN22: genetic organization and correlation with genetic properties. *Plasmid*, 32(1), 55–69.

[Kaysser *et al.*, 2009] Kaysser, L., Lutsch, L., Siebenberg, S., Wemakor, E., Kammerer, B., & Gust, B. (2009). Identification and manipulation of the caprazamycin gene cluster lead to new simplified liponucleoside antibiotics and give insights into the biosynthetic pathway. *The Journal of Biological Chemistry*, 284(22), 14987–14996. PMID: 19351877 PMCID: 2685681.

[Keiler, 2008] Keiler, K. C. (2008). Biology of trans-translation. *Annual Review of Microbiology*, 62, 133–151. PMID: 18557701.

[Kin *et al.*, 2007] Kin, T., Yamada, K., Terai, G., Okida, H., Yoshinari, Y., Ono, Y., Kojima, A., Kimura, Y., Komori, T., & Asai, K. (2007). fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences. *Nucleic Acids Research*, 35(Database issue), D145–148. PMID: 17099231.

[Kingsford *et al.*, 2007a] Kingsford, C. L., Ayanbule, K., & Salzberg, S. L. (2007a). Rapid, accurate, computational discovery of rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biology*, 8(2), R22.

[Kingsford *et al.*, 2007b] Kingsford, C. L., Ayanbule, K., & Salzberg, S. L. (2007b). Rapid, accurate, computational discovery of rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biology*, 8(2), R22.

[Kirillov *et al.*, 1997] Kirillov, S., Vitali, L. A., Goldstein, B. P., Monti, F., Semenkov, Y., Makhno, V., Ripa, S., Pon, C. L., & Gualerzi, C. O. (1997). Purpuromycin: an antibiotic inhibiting tRNA aminoacylation. *RNA (New York, N.Y.)*, 3(8), 905–913. PMID: 9257649.

[Knippers, 2001] Knippers, R. (2001). *Molekulare Genetik*. Thieme, Stuttgart, 8 edition.

[Kotowska *et al.*, 2002] Kotowska, M., Pawlik, K., Butler, A. R., Cundliffe, E., Takano, E., & Kuczek, K. (2002). Type II thioesterase from *Streptomyces coelicolor* A3(2). *Microbiology*, 148(6), 1777 –1783.

[Kranz & Gennis, 1985] Kranz, R. G. & Gennis, R. B. (1985). Immunological investigation of the distribution of cytochromes related to the two terminal oxidases of *Escherichia coli* in other Gram-negative bacteria. *Journal of Bacteriology*, 161(2), 709–713. PMID: 2981822.

[Krause *et al.*, 2007] Krause, L., McHardy, A. C., Nattkemper, T. W., Pühler, A., Stoye, J., & Meyer, F. (2007). GISMO–gene identification using a support vector machine for ORF classification. *Nucleic Acids Research*, 35(2), 540–549. PMID: 17175534.

[Krogh *et al.*, 2001] Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology*, 305(3), 567–580. PMID: 11152613.

[Krätzschmar *et al.*, 1989] Krätzschmar, J., Krause, M., & Marahiel, M. A. (1989). Gramicidin S biosynthesis operon containing the structural genes *grsA* and *grsB* has an open reading frame encoding a protein homologous to fatty acid thioesterases. *Journal of Bacteriology*, 171(10), 5422–5429. PMID: 2477357 PMCID: 210379.

[Kryukov & Gladyshev, 2004] Kryukov, G. V. & Gladyshev, V. N. (2004). The prokaryotic selenoproteome. *EMBO Rep*, 5(5), 538–543.

[Kuhstoss *et al.*, 1991] Kuhstoss, S., Richardson, M. A., & Rao, R. N. (1991). Plasmid cloning vectors that integrate site-specifically in *Streptomyces* spp. *Gene*, 97(1), 143–146. PMID: 1995427.

[Kushner, 2002] Kushner, S. R. (2002). mRNA decay in *Escherichia coli* comes of age. *Journal of Bacteriology*, 184(17), 4658 –4665.

[Łabaj *et al.*, 2011] Łabaj, P. P., Leparc, G. G., Linggi, B. E., Markillie, L. M., Wiley, H. S., & Kreil, D. P. (2011). Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. *Bioinformatics (Oxford, England)*, 27(13), i383–391. PMID: 21685096.

[Lagesen *et al.*, 2007] Lagesen, K., Hallin, P., Rødland, E. A., Stæfeldt, H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, 35(9), 3100 –3108.

[Langmead *et al.*, 2009] Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3), R25. PMID: 19261174.

[Latreille *et al.*, 2007] Latreille, P., Norton, S., Goldman, B. S., Henkhaus, J., Miller, N., Barbazuk, B., Bode, H. B., Darby, C., Du, Z., Forst, S., Gaudriault, S., Goodner, B., Goodrich-Blair, H., & Slater, S. (2007). Optical mapping as a routine tool for bacterial genome sequence finishing. *BMC Genomics*, 8, 321. PMID: 17868451.

[Lazzarini *et al.*, 2001] Lazzarini, A., Cavaletti, L., Toppo, G., & Marinelli, F. (2001). Rare genera of actinomycetes as potential producers of new antibiotics. *Antonie Van Leeuwenhoek*, 79(3-4), 399–405. PMID: 11816986.

[Lechevalier & Lechevalier, 1970] Lechevalier, H. & Lechevalier, M. (1970). In *The Actinomycetales* (pp. 393– 405). Jena: Veb G. Fisher.

[Lee *et al.*, 2008] Lee, J., Hai, T., Pape, H., Kim, T., & Suh, J. (2008). Three trehalose synthetic pathways in the acarbose-producing *Actinoplanes* sp. SN223/29 and evidence for the TreY role in biosynthesis of component C. *Applied Microbiology and Biotechnology*, 80(5), 767–778. PMID: 18663442.

[Lee *et al.*, 1997] Lee, S., Sauerbrei, B., Niggemann, J., & Egelkrout, E. (1997). Biosynthetic studies on the alpha-glucosidase inhibitor acarbose in *Actinoplanes* sp.: source of the maltose unit. *The Journal of Antibiotics*, 50(11), 954–960. PMID: 9592570.

[Li & Durbin, 2009] Li, H. & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 25(14), 1754–1760. PMID: 19451168.

[Li *et al.*, 2008] Li, R., Li, Y., Kristiansen, K., & Wang, J. (2008). SOAP: short oligonucleotide alignment program. *Bioinformatics (Oxford, England)*, 24(5), 713–714. PMID: 18227114.

[Li *et al.*, 2006] Li, W., Xin, Y., McNeil, M. R., & Ma, Y. (2006). *rmlB* and *rmlC* genes are essential for growth of mycobacteria. *Biochemical and Biophysical Research Communications*, 342(1), 170–178.

[Licht *et al.*, 2011] Licht, A., Bulut, H., Scheffel, F., Daumke, O., Wehmeier, U. F., Saenger, W., Schneider, E., & Vahedi-Faridi, A. (2011). Crystal structures of the bacterial solute receptor AcbH displaying an exclusive substrate preference for ß-d-Galactopyranose. *Journal of Molecular Biology*, 406(1), 92–105. PMID: 21168419.

[Linke *et al.*, 2006] Linke, B., McHardy, A. C., Neuweger, H., Krause, L., & Meyer, F. (2006). REGANOR: a gene prediction server for prokaryotic genomes and a database of high quality gene predictions for prokaryotes. *Applied Bioinformatics*, 5(3), 193–198. PMID: 16922601.

[Lister *et al.*, 2008] Lister, R., O'Malley, R. C., Tonti-Filippini, J., Gregory, B. D., Berry, C. C., Millar, A. H., & Ecker, J. R. (2008). Highly integrated single-base

resolution maps of the epigenome in arabidopsis. *Cell*, 133(3), 523–536. PMID: 18423832.

[Liu, 2004] Liu, C. (2004). NONCODE: an integrated knowledge database of non-coding RNAs. *Nucleic Acids Research*, 33(Database issue), D112–D115.

[Liu & Thorson, 1994] Liu, H. W. & Thorson, J. S. (1994). Pathways and mechanisms in the biogenesis of novel deoxysugars by bacteria. *Annual Review of Microbiology*, 48, 223–256. PMID: 7826006.

[Loewen & Switala, 1995] Loewen, P. C. & Switala, J. (1995). Template secondary structure can increase the error frequency of the DNA polymerase from thermus aquaticus. *Gene*, 164(1), 59–63.

[Lowe & Eddy, 1997] Lowe, T. M. & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, 25(5), 955–964. PMID: 9023104.

[Lu *et al.*, 2003] Lu, Z. H., Dameron, C. T., & Solioz, M. (2003). The *Enterococcus hirae* paradigm of copper homeostasis: copper chaperone turnover, interactions, and transactions. *Biometals: An International Journal on the Role of Metal Ions in Biology, Biochemistry, and Medicine*, 16(1), 137–143. PMID: 12572673.

[Magnani & Solioz, 2005] Magnani, D. & Solioz, M. (2005). Copper chaperone cycling and degradation in the regulation of the Cop operon of *Enterococcus hirae*. *BioMetals*, 18, 407–412.

[Mahmud *et al.*, 1999] Mahmud, T., Tornus, I., Egelkrout, E., Wolf, E., Uy, C., Floss, H. G., & Lee, S. (1999). Biosynthetic studies on the α-Glucosidase inhibitor acarbose in *Actinoplanes* sp.: 2-epi-5-epi-valiolone is the direct precursor of the valienamine moiety. *J. Am. Chem. Soc.*, 121(30), 6973–6983.

[Marchler-Bauer *et al.*, 2011a] Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., Fong, J. H., Geer, L. Y., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Jackson, J. D., Ke, Z., Lanczycki, C. J., Lu, F., Marchler, G. H., Mullokandov, M., Omelchenko, M. V., Robertson, C. L., Song, J. S., Thanki, N., Yamashita, R. A., Zhang, D., Zhang, N., Zheng, C., & Bryant, S. H. (2011a). CDD: a conserved domain database for the functional annotation of proteins. *Nucleic Acids Research*, 39(Database issue), D225–229. PMID: 21109532.

[Marchler-Bauer *et al.*, 2011b] Marchler-Bauer, A., Lu, S., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., Fong, J. H., Geer, L. Y., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Jackson, J. D., Ke, Z., Lanczycki, C. J., Lu, F., Marchler, G. H., Mullokandov, M., Omelchenko, M. V., Robertson, C. L., Song, J. S., Thanki, N., Yamashita, R. A., Zhang, D., Zhang, N., Zheng, C., &

Bryant, S. H. (2011b). CDD: a conserved domain database for the functional annotation of proteins. *Nucleic Acids Research*, 39(Database issue), D225–229. PMID: 21109532.

[Marchler-Bauer *et al.*, 2002] Marchler-Bauer, A., Panchenko, A. R., Shoemaker, B. A., Thiessen, P. A., Geer, L. Y., & Bryant, S. H. (2002). CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Research*, 30(1), 281–283. PMID: 11752315.

[Mardis, 2008a] Mardis, E. R. (2008a). The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, 24(3), 133–141.

[Mardis, 2008b] Mardis, E. R. (2008b). Next-generation DNA sequencing methods. *Annual Review of Genomics and Human Genetics*, 9, 387–402. PMID: 18576944.

[Margulies *et al.*, 2005] Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., & Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057), 376–380. PMID: 16056220.

[Massé *et al.*, 2003] Massé, E., Majdalani, N., & Gottesman, S. (2003). Regulatory roles for small RNAs in bacteria. *Current Opinion in Microbiology*, 6(2), 120–124. PMID: 12732300.

[Massa *et al.*, 2011] Massa, A. N., Childs, K. L., Lin, H., Bryan, G. J., Giuliano, G., & Buell, C. R. (2011). The transcriptome of the reference potato genome *Solanum tuberosum* group Phureja clone DM1-3 516R44. *PloS One*, 6(10), e26801. PMID: 22046362.

[Mast *et al.*, 2011] Mast, Y., Weber, T., Gölz, M., Ort-Winklbauer, R., Gondran, A., Wohlleben, W., & Schinko, E. (2011). Characterization of the 'pristinamycin supercluster' of *Streptomyces pristinaespiralis*. *Microbial Biotechnology*, 4(2), 192–206.

[Mathé *et al.*, 2002] Mathé, C., Sagot, M., Schiex, T., & Rouzé, P. (2002). Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Research*, 30(19), 4103–4117.

[Mathias, 1995] Mathias, L. (1995). Cytochromes of archaeal electron transfer chains. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1229(1), 1–22.

[Maupin-Furlow *et al.*, 1995] Maupin-Furlow, J. A., Rosentel, J. K., Lee, J. H., Deppenmeier, U., Gunsalus, R. P., & Shanmugam, K. T. (1995). Genetic analysis of the *modABCD* (molybdate transport) operon of *Escherichia coli*. *Journal of Bacteriology*, 177(17), 4851–4856. PMID: 7665460 PMCID: 177257.

[Maxam & Gilbert, 1977] Maxam, A. M. & Gilbert, W. (1977). A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 74(2), 560–564. PMID: 265521.

[McDowell *et al.*, 1998] McDowell, D. G., Burns, N. A., & Parkes, H. C. (1998). Localised sequence regions possessing high melting temperatures prevent the amplification of a DNA mimic in competitive PCR. *Nucleic Acids Research*, 26(14), 3340–3347. PMID: 9649616.

[McHardy *et al.*, 2004] McHardy, A. C., Goesmann, A., Pühler, A., & Meyer, F. (2004). Development of joint application strategies for two microbial gene finders. *Bioinformatics (Oxford, England)*, 20(10), 1622–1631. PMID: 14988122.

[Medema *et al.*, 2011] Medema, M. H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M. A., Weber, T., Takano, E., & Breitling, R. (2011). antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Research*.

[Medema *et al.*, 2010] Medema, M. H., Trefzer, A., Kovalchuk, A., van den Berg, M., Müller, U., Heijne, W., Wu, L., Alam, M. T., Ronning, C. M., Nierman, W. C., Bovenberg, R. A. L., Breitling, R., & Takano, E. (2010). The sequence of a 1.8-Mb bacterial linear plasmid reveals a rich evolutionary reservoir of secondary metabolic pathways. 2, 212–224. PMID: 20624727 PMCID: 2997539.

[Mehling *et al.*, 1995a] Mehling, A., Wehmeier, U. F., & Piepersberg, W. (1995a). Application of random amplified polymorphic DNA (RAPD) assays in identifying conserved regions of actinomycete genomes. *FEMS Microbiology Letters*, 128(2), 119–125. PMID: 7750729.

[Mehling *et al.*, 1995b] Mehling, A., Wehmeier, U. F., & Piepersberg, W. (1995b). Nucleotide sequences of *Streptomycete* 16S ribosomal DNA: towards a specific identification system for *Streptomycetes* using PCR. *Microbiology (Reading, England)*, 141 (Pt 9), 2139–2147. PMID: 7496525.

[Meier & Burkart, 2009] Meier, J. L. & Burkart, M. D. (2009). The chemical biology of modular biosynthetic enzymes. *Chemical Society Reviews*, 38(7), 2012–2045. PMID: 19551180.

[Melançon & Liu, 2007] Melançon, Charles E, r. & Liu, H. (2007). Engineered biosynthesis of macrolide derivatives bearing the non-natural deoxysugars 4-epi-D-mycaminose and 3-n-monomethylamino-3-deoxy-D-fucose. *Journal of the American Chemical Society*, 129(16), 4896–4897. PMID: 17388593.

[Meyer *et al.*, 2003] Meyer, F., Goesmann, A., McHardy, A. C., Bartels, D., Bekel, T., Clausen, J., Kalinowski, J., Linke, B., Rupp, O., Giegerich, R., & Pühler, A. (2003). GenDB–an open source genome annotation system for prokaryote genomes. *Nucleic Acids Research*, 31(8), 2187–2195. PMID: 12682369.

[Mitra *et al.*, 2009] Mitra, A., Angamuthu, K., Jayashree, H. V., & Nagaraja, V. (2009). Occurrence, divergence and evolution of intrinsic terminators across eubacteria. *Genomics*, 94(2), 110–116. PMID: 19393739.

[Mituyama *et al.*, 2009] Mituyama, T., Yamada, K., Hattori, E., Okida, H., Ono, Y., Terai, G., Yoshizawa, A., Komori, T., & Asai, K. (2009). The functional RNA database 3.0: databases to support mining and annotation of functional RNAs. *Nucleic Acids Research*, 37(Database issue), D89–92. PMID: 18948287.

[Müller *et al.*, 1980] Müller, L., Junge, B., Frommer, W., Schmidet, D., & Truscheit, E. (1980). Acarbose (Bay g5421) and homologous α-glucosidase inhibitors from actinoplanaceae. In U. Brodbeck (Ed.), *Enzyme inhibitors*. Weinheim: Verlag Chemie.

[Mootz & Marahiel, 1997] Mootz, H. D. & Marahiel, M. A. (1997). The tyrocidine biosynthesis operon of *Bacillus brevis*: complete nucleotide sequence and biochemical characterization of functional internal adenylation domains. *Journal of Bacteriology*, 179(21), 6843–6850. PMID: 9352938.

[Mortazavi *et al.*, 2008] Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, 5(7), 621–628. PMID: 18516045.

[Munk *et al.*, 2009] Munk, C., Lapidus, A., Copeland, A., Jando, M., Mayilraj, S., Rio, T. G. D., Nolan, M., Chen, F., Lucas, S., Tice, H., Cheng, J., Han, C., Detter, J. C., Bruce, D., Goodwin, L., Chain, P., Pitluck, S., Goker, M., Ovchinnikova, G., Pati, A., Ivanova, N., Mavromatis, K., Chen, A., Palaniappan, K., Land, M., Hauser, L., Chang, Y., Jeffries, C. D., Bristow, J., Eisen, J. A., Markowitz, V., Hugenholtz, P., Kyrpides, N. C., & Klenk, H. (2009). Complete genome sequence of *Stackebrandtia nassauensis* type strain (LLR-40K-21^T). *Standards in Genomic Sciences*, 1(3), 292–299.

[Nagalakshmi *et al.*, 2008] Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., & Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science (New York, N.Y.)*, 320(5881), 1344–1349. PMID: 18451266.

[Nahoum *et al.*, 2000] Nahoum, V., Roux, G., Anton, V., Rougé, P., Puigserver, A., Bischoff, H., Henrissat, B., & Payan, F. (2000). Crystal structures of human pancreatic alpha-amylase in complex with carbohydrate and proteinaceous inhibitors. *The Biochemical Journal*, 346 Pt 1, 201–208. PMID: 10657258.

[Nahvi *et al.*, 2002] Nahvi, A., Sudarsan, N., Ebert, M. S., Zou, X., Brown, K. L., & Breaker, R. R. (2002). Genetic control by a metabolite binding mRNA. *Chemistry & Biology*, 9(9), 1043. PMID: 12323379.

[Naville & Gautheret, 2009] Naville, M. & Gautheret, D. (2009). Transcription attenuation in bacteria: theme and variations. *Briefings in Functional Genomics & Proteomics*, 8(6), 482–492. PMID: 19651704.

[NC-ICBMB & Webb, 1992] NC-ICBMB & Webb, E. C. (1992). *Enzyme Nomenclature 1992: Recommendations of the NCIUBMB on the Nomenclature and Classification of Enzymes*. San Diego, California: Academic Press, 1 edition.

[Niedringhaus *et al.*, 2011] Niedringhaus, T. P., Milanova, D., Kerby, M. B., Snyder, M. P., & Barron, A. E. (2011). Landscape of Next-Generation sequencing technologies. *Anal. Chem.*, 83(12), 4327–4341.

[Nielsen *et al.*, 1997] Nielsen, H., Engelbrecht, J., Brunak, S., & von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering*, 10(1), 1–6. PMID: 9051728.

[Nielsen & Krogh, 1998] Nielsen, H. & Krogh, A. (1998). Prediction of signal peptides and signal anchors by a hidden Markov model. *Proc Int Conf Intell Syst Mol Biol*, 6, 122–130. PMID: 9783217.

[Oh *et al.*, 2007] Oh, T., Mo, S. J., Yoon, Y. J., & Sohng, J. K. (2007). Discovery and molecular engineering of sugar-containing natural product biosynthetic pathways in actinomycetes. *Journal of Microbiology and Biotechnology*, 17(12), 1909–1921. PMID: 18167436.

[Olano *et al.*, 2008] Olano, C., Lombó, F., Méndez, C., & Salas, J. A. (2008). Improving production of bioactive secondary metabolites in actinomycetes by metabolic engineering. *Metabolic Engineering*, 10(5), 281–292. PMID: 18674632.

[Ollinger *et al.*, 2006] Ollinger, J., Song, K., Antelmann, H., Hecker, M., & Helmann, J. D. (2006). Role of the fur regulon in iron transport in *Bacillus subtilis*. *Journal of Bacteriology*, 188(10), 3664–3673. PMID: 16672620.

[Omer *et al.*, 1988] Omer, C. A., Stein, D., & Cohen, S. N. (1988). Site-specific insertion of biologically functional adventitious genes into the *Streptomyces lividans* chromosome. *Journal of Bacteriology*, 170(5), 2174–2184. PMID: 2834330.

[Oshlack *et al.*, 2010] Oshlack, A., Robinson, M. D., & Young, M. D. (2010). From RNA-seq reads to differential expression results. *Genome Biology*, 11(12), 220. PMID: 21176179.

[Ozsolak & Milos, 2011] Ozsolak, F. & Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nature reviews. Genetics*, 12(2), 87–98. PMID: 21191423 PMCID: 3031867.

[Ozsolak *et al.*, 2009] Ozsolak, F., Platt, A. R., Jones, D. R., Reifenberger, J. G., Sass, L. E., McInerney, P., Thompson, J. F., Bowers, J., Jarosz, M., & Milos, P. M. (2009). Direct RNA sequencing. *Nature*, 461(7265), 814–818. PMID: 19776739.

[Padmanabhan *et al.*, 2001] Padmanabhan, S., Elías-Arnanz, M., Carpio, E., Aparicio, P., & Murillo, F. J. (2001). Domain architecture of a high mobility group a-type bacterial transcriptional factor. *Journal of Biological Chemistry*, 276(45), 41566 – 41575.

[Pang *et al.*, 2006] Pang, K. C., Frith, M. C., & Mattick, J. S. (2006). Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends in Genetics: TIG*, 22(1), 1–5. PMID: 16290135.

[Parenti & Coronelli, 1979] Parenti, F. & Coronelli, C. (1979). Members of the genus *Actinoplanes* and their antibiotics. *Annual Review of Microbiology*, 33(1), 389–411.

[Parenti *et al.*, 1975] Parenti, F., Pagani, H., & Beretta, G. (1975). Lipiarmycin, a new antibiotic from *Actinoplanes*. I. description of the producer strain and fermentation studies. *The Journal of Antibiotics*, 28(4), 247–252. PMID: 1150527.

[Patschkowski *et al.*, 2000] Patschkowski, T., Bates, D., & Kiley, P. (2000). Mechanisms for sensing and responding to oxygen deprivation. In *Bacterial Stress Responses* (pp. 61–78). Washington, DC: ASM Press.

[Piepersberg, 1993] Piepersberg, W. (1993). *Streptotmyces* and *Corynebacteria*. In P. A. S. P. Rehm H-J, Reed G (Ed.), *Biotechnology* (pp. 434–468). Wiley-VCH Verlag GmbH.

[Piepersberg *et al.*, 2002] Piepersberg, W., Diaz-Guardamino Uribe, P., Stratmann, A., Thomas, H., Wehmeier, U., & Zhang, C. (2002). Developments in the biosynthesis and regulation of aminoglycosides. In *Microbial Secondary Metabolites: Biosynthesis, Genetics and Regulation* (pp. 1–26). Kerala, India: Research Signpost. edited by Francisco Fierro and Juan Francisco Martín.

[Piepersberg & Distler, 1997] Piepersberg, W. & Distler, J. (1997). Aminoglycosides and sugar components in other secondary metabolites. In P. A. S. P. Rehm H-J, Reed G (Ed.), *Biotechnology (Products of secondary metabolism, vol 7)* (pp. 397–488). Wiley-VCH Verlag GmbH.

[Puri-Taneja *et al.*, 2007] Puri-Taneja, A., Schau, M., Chen, Y., & Hulett, F. M. (2007). Regulators of the *Bacillus subtilis cydABCD* operon: Identification of a negative regulator, CcpA, and a positive regulator, ResD. *Journal of Bacteriology*, 189(9), 3348–3358. PMID: 17322317 PMCID: 1855890.

[Quail *et al.*, 1996] Quail, M. A., Jordan, P., Grogan, J. M., Butt, J. N., Lutz, M., Thomson, A. J., Andrews, S. C., & Guest, J. R. (1996). Spectroscopic and

voltammetric characterisation of the bacterioferritin-associated ferredoxin of *Escherichia coli*. *Biochemical and Biophysical Research Communications*, 229(2), 635–642. PMID: 8954950.

[Rajasekaran *et al.*, 2010] Rajasekaran, M. B., Nilapwar, S., Andrews, S. C., & Watson, K. A. (2010). EfeO-cupredoxins: major new members of the cupredoxin superfamily with roles in bacterial iron transport. *Biometals: An International Journal on the Role of Metal Ions in Biology, Biochemistry, and Medicine*, 23(1), 1–17. PMID: 19701722.

[Ramos *et al.*, 2005] Ramos, J. L., Martínez-Bueno, M., Molina-Henares, A. J., Terán, W., Watanabe, K., Zhang, X., Gallegos, M. T., Brennan, R., & Tobes, R. (2005). The TetR family of transcriptional repressors. *Microbiology and Molecular Biology Reviews*, 69(2), 326–356. PMID: 15944459 PMCID: 1197418.

[Rauschenbusch & Schmidt, 1978] Rauschenbusch, E. & Schmidt, D. (1978). Verfahren zur Isolierung von (O4,6-Dideoxy-4[[1S-(1,4,6/5)-4,5,6-trihydroxy-3-hydroxymethyl-2-cyclohexen-1-yl]-amino]-a-D-glucopyranosyl-(1→4)-O-a-D-glucopyranosyl-(1→4)-D-glucopyranose) aus Kulturbrühen. Patent. DE 2719912 (US patent 4,174,439).

[Raynal *et al.*, 2002] Raynal, A., Friedmann, A., Tuphile, K., Guerineau, M., & Pernodet, J. (2002). Characterization of the attP site of the integrative element pSAM2 from *Streptomyces ambofaciens*. *Microbiology (Reading, England)*, 148(Pt 1), 61–67. PMID: 11782499.

[Reichard, 1993] Reichard, P. (1993). From RNA to DNA, why so many ribonucleotide reductases? *Science*, 260(5115), 1773 –1777.

[Retief, 2000] Retief, J. D. (2000). Phylogenetic analysis using PHYLIP. *Methods in Molecular Biology (Clifton, N.J.)*, 132, 243–258. PMID: 10547839.

[Robinson *et al.*, 2010] Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, 26(1), 139–140. PMID: 19910308.

[Rockser & Wehmeier, 2009] Rockser, Y. & Wehmeier, U. F. (2009). The *gac*-gene cluster for the production of acarbose from *Streptomyces glaucescens* GLA.O: identification, isolation and characterization. *Journal of Biotechnology*, 140(1-2), 114–123. PMID: 19059289.

[Roglic & Unwin, 2010] Roglic, G. & Unwin, N. (2010). Mortality attributable to diabetes: estimates for the year 2010. *Diabetes Research and Clinical Practice*, 87(1), 15–19. PMID: 19914728.

[Roh *et al.*, 2011] Roh, H., Uguru, G. C., Ko, H., Kim, S., Kim, B., Goodfellow, M., Bull, A. T., Kim, K. H., Bibb, M. J., Choi, I., & Stach, J. E. M. (2011). Genome

sequence of the abyssomicin- and proximicin-producing marine actinomycete *Verrucosispora maris* AB-18-032. *Journal of Bacteriology*, 193(13), 3391–3392. PMID: 21551311.

[Ruan & Ruan, 2012] Ruan, X. & Ruan, Y. (2012). Genome wide full-length transcript analysis using 5' and 3' paired-end-tag next generation sequencing (RNA-PET). *Methods in Molecular Biology (Clifton, N.J.)*, 809, 535–562. PMID: 22113299.

[Rutherford *et al.*, 2000] Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M., & Barrell, B. (2000). Artemis: sequence visualization and annotation. *Bioinformatics*, 16(10), 944–945.

[Sahdev *et al.*, 2007] Sahdev, S., Saini, S., Tiwari, P., Saxena, S., & Singh Saini, K. (2007). Amplification of GC-rich genes by following a combination strategy of primer design, enhancers and modified PCR cycle conditions. *Molecular and Cellular Probes*, 21(4), 303–307. PMID: 17490855.

[Saitou & Nei, 1987] Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4), 406–425. PMID: 3447015.

[Salgado *et al.*, 2000] Salgado, H., Moreno-Hagelsieb, G., Smith, T. F., & Collado-Vides, J. (2000). Operons in *Escherichia coli*: genomic analyses and predictions. *Proceedings of the National Academy of Sciences of the United States of America*, 97(12), 6652–6657. PMID: 10823905.

[Salzberg *et al.*, 1998] Salzberg, S. L., Delcher, A. L., Kasif, S., & White, O. (1998). Microbial gene identification using interpolated Markov models. *Nucleic Acids Research*, 26(2), 544–548. PMID: 9421513.

[Samulitis *et al.*, 1987] Samulitis, B. K., Goda, T., Lee, S. M., & Koldovský, O. (1987). Inhibitory mechanism of acarbose and 1-deoxynojirimycin derivatives on carbohydrases in rat small intestine. *Drugs Under Experimental and Clinical Research*, 13(8), 517–524. PMID: 2962844.

[Sanger & Coulson, 1975] Sanger, F. & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94(3), 441–448. PMID: 1100841.

[Schau *et al.*, 2004] Schau, M., Chen, Y., & Hulett, F. M. (2004). *Bacillus subtilis* YdiH is a direct negative regulator of the *cydABCD* operon. *Journal of Bacteriology*, 186(14), 4585–4595. PMID: 15231791.

[Schauder & Bassler, 2001] Schauder, S. & Bassler, B. L. (2001). The languages of bacteria. *Genes & Development*, 15(12), 1468 –1480.

[Schedel, 2006] Schedel, M. (2006). Weiße Biotechnologie bei Bayer HealthCare Product Supply: Mehr als 30 Jahre Erfahrung. *Chemie Ingenieur Technik*, 78(4), 485–489.

[Schmidt *et al.*, 1977] Schmidt, D. D., Frommer, W., Junge, B., Müller, L., Wingender, W., Truscheit, E., & Schäfer, D. (1977). Alpha-glucosidase inhibitors. new complex oligosaccharides of microbial origin. *Die Naturwissenschaften*, 64(10), 535–536. PMID: 337162.

[Schwientek *et al.*, 2012] Schwientek, P., Szczepanowski, R., Rückert, C., Kalinowski, J., Klein, A., Selber, K., Wehmeier, U. F., Stoye, J., & Pühler, A. (2012). The complete genome sequence of the acarbose producer *Actinoplanes* sp. SE50/110. *BMC Genomics*. Submitted.

[Sezonov *et al.*, 1995] Sezonov, G., Hagège, J., Pernodet, J. L., Friedmann, A., & Guérineau, M. (1995). Characterization of *pra*, a gene for replication control in pSAM2, the integrating element of *Streptomyces ambofaciens*. *Molecular Microbiology*, 17(3), 533–544. PMID: 8559072.

[Sharma *et al.*, 2010] Sharma, C. M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiss, S., Sittka, A., Chabas, S., Reiche, K., Hackermüller, J., Reinhardt, R., Stadler, P. F., & Vogel, J. (2010). The primary transcriptome of the major human pathogen helicobacter pylori. *Nature*, 464(7286), 250–255. PMID: 20164839.

[Shen *et al.*, 2001] Shen, B., Du, L., Sánchez, C., Edwards, D. J., Chen, M., & Murrell, J. M. (2001). The biosynthetic gene cluster for the anticancer drug bleomycin from *Streptomyces verticillus* ATCC15003 as a model for hybrid peptide-polyketide natural product biosynthesis. *Journal of Industrial Microbiology and Biotechnology*, 27(6), 378–385.

[Sigrist *et al.*, 1975] Sigrist, H., Ronner, P., & Semenza, G. (1975). A hydrophobic form of the small-intestinal sucrase-isomaltase complex. *Biochimica Et Biophysica Acta*, 406(3), 433–446. PMID: 1182172.

[Simpson *et al.*, 2009] Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., & Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome Research*, 19(6), 1117–1123. PMID: 19251739.

[Singh *et al.*, 2007] Singh, S. B., Occi, J., Jayasuriya, H., Herath, K., Motyl, M., Dorso, K., Gill, C., Hickey, E., Overbye, K. M., Barrett, J. F., & Masurekar, P. (2007). Antibacterial evaluations of thiazomycin- a potent thiazolyl peptide antibiotic from amycolatopsis fastidiosa. *The Journal of Antibiotics*, 60(9), 565–571. PMID: 17917239.

[Singh *et al.*, 2011] Singh, S. S., Typas, A., Hengge, R., & Grainger, D. C. (2011). *Escherichia coli* σ70 senses sequence and conformation of the promoter spacer region. *Nucleic Acids Research*, 39(12), 5109–5118. PMID: 21398630 PMCID: 3130263.

[Smokvina *et al.*, 1990] Smokvina, T., Mazodier, P., Boccard, F., Thompson, C. J., & Guérineau, M. (1990). Construction of a series of pSAM2-based integrative vectors for use in actinomycetes. *Gene*, 94(1), 53–59. PMID: 2227452.

[Sonnhammer *et al.*, 1998] Sonnhammer, E. L., von Heijne, G., & Krogh, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol*, 6, 175–182. PMID: 9783223.

[Spiess *et al.*, 2004] Spiess, A., Mueller, N., & Ivell, R. (2004). Trehalose is a potent PCR enhancer: Lowering of DNA melting temperature and thermal stabilization of taq polymerase by the disaccharide trehalose,. *Clin Chem*, 50(7), 1256–1259.

[Staden, 1979] Staden, R. (1979). A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research*, 6(7), 2601–2610. PMID: 461197.

[Stockmann & Piepersberg, 1992] Stockmann, M. & Piepersberg, W. (1992). Gene probes for the detection of 6-deoxyhexose metabolism in secondary metabolite-producing *Streptomycetes*. *FEMS Microbiology Letters*, 69(2), 185–189. PMID: 1537548.

[Stratmann, 1997] Stratmann, A. (1997). *Identifizierung eines Acarbose-Biosynthesegenclusters in Actinoplanes sp. und Charakterisierung ausgewählter Enzyme des Acarbose-Stoffwechsels*. PhD thesis, Bergische Universität, Wuppertal.

[Stratmann *et al.*, 1999] Stratmann, A., Mahmud, T., Lee, S., Distler, J., Floss, H. G., & Piepersberg, W. (1999). The AcbC protein from *Actinoplanes* species is a C7-cyclitol synthase related to 3-dehydroquinate synthases and is involved in the biosynthesis of the α-glucosidase inhibitor acarbose. *Journal of Biological Chemistry*, 274(16), 10889 –10896.

[Sun *et al.*, 2006] Sun, L., Campbell, F. E., Zahler, N. H., & Harris, M. E. (2006). Evidence that substrate-specific effects of C5 protein lead to uniformity in binding and catalysis by RNase P. *The EMBO Journal*, 25(17), 3998–4007. PMID: 16932744 PMCID: 1560353.

[Šuput *et al.*, 1967] Šuput, J., Lechevalier, M. P., & Lechevalier, H. A. (1967). Chemical composition of variants of aerobic actinomycetes. *Applied Microbiology*, 15(6), 1356–1361. PMID: 16349745 PMCID: 547199.

[Tamura *et al.*, 2007] Tamura, K., Dudley, J., Nei, M., & Kumar, S. (2007). MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular Biology and Evolution*, 24(8), 1596–1599. PMID: 17488738.

[Tarazona *et al.*, 2011] Tarazona, S., García-Alcalde, F., Dopazo, J., Ferrer, A., & Conesa, A. (2011). Differential expression in RNA-seq: a matter of depth. *Genome Research*. PMID: 21903743.

[Tatusov *et al.*, 2003] Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., & Natale, D. A. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4, 41. PMID: 12969510.

[Tatusov *et al.*, 1997] Tatusov, R. L., Koonin, E. V., & Lipman, D. J. (1997). A genomic perspective on protein families. *Science (New York, N.Y.)*, 278(5338), 631–637. PMID: 9381173.

[Tatusov *et al.*, 2001] Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D., & Koonin, E. V. (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Research*, 29(1), 22–28. PMID: 11125040.

[Tauch *et al.*, 2008] Tauch, A., Trost, E., Tilker, A., Ludewig, U., Schneiker, S., Goesmann, A., Arnold, W., Bekel, T., Brinkrolf, K., Brune, I., Götker, S., Kalinowski, J., Kamp, P., Lobo, F. P., Viehoever, P., Weisshaar, B., Soriano, F., Dröge, M., & Pühler, A. (2008). The lifestyle of *Corynebacterium urealyticum* derived from its complete genome sequence established by pyrosequencing. *Journal of Biotechnology*, 136(1-2), 11–21.

[te Poele *et al.*, 2008] te Poele, E. M., Bolhuis, H., & Dijkhuizen, L. (2008). Actinomycete integrative and conjugative elements. *Antonie Van Leeuwenhoek*, 94(1), 127–143. PMID: 18523858.

[Thomas, 2001] Thomas, H. (2001). *Acarbose-Metabolismus in Actinoplanes sp. SE50/110*. PhD thesis, Wuppertal University, Wuppertal.

[Thompson *et al.*, 2002] Thompson, J. D., Gibson, T. J., & Higgins, D. G. (2002). Multiple sequence alignment using ClustalW and ClustalX. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]*, Chapter 2, Unit 2.3. PMID: 18792934.

[Thompson & Milos, 2011] Thompson, J. F. & Milos, P. M. (2011). The properties and applications of single-molecule DNA sequencing. *Genome Biology*, 12(2), 217.

[Torrents *et al.*, 2002] Torrents, E., Aloy, P., Gibert, I., & Rodríguez-Trelles, F. (2002). Ribonucleotide reductases: Divergent evolution of an ancient enzyme. *Journal of Molecular Evolution*, 55(2), 138–152.

[Trapnell *et al.*, 2010] Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotech*, 28(5), 511–515.

[Trost *et al.*, 2010] Trost, E., Götker, S., Schneider, J., Schneiker-Bekel, S., Szczepanowski, R., Tilker, A., Viehoever, P., Arnold, W., Bekel, T., Blom, J., Gartemann, K., Linke, B., Goesmann, A., Pühler, A., Shukla, S. K., & Tauch, A. (2010). Complete genome sequence and lifestyle of black-pigmented *Corynebacterium aurimucosum* ATCC 700975 (formerly *C. nigricans* CN-1) isolated from a vaginal swab of a woman with spontaneous abortion. *BMC Genomics*, 11, 91. PMID: 20137072.

[Truscheit *et al.*, 1981] Truscheit, E., Frommer, W., Junge, B., Müller, L., Schmidt, D. D., & Wingender, W. (1981). Chemistry and biochemistry of microbial α-glucosidase inhibitors. *Angewandte Chemie International Edition in English*, 20(9), 744–761.

[Truscheit *et al.*, 1988] Truscheit, E., Junge, B., Müller, L., Puls, W., & Schmidt, D. (1988). Microbial alpha-glucosidase inhibitors: chemistry, biochemistry and therapeutic potential. *Prog Clin Biochem Med*, (7), 17.

[Turner & Jenkins, 1995] Turner, S. L. & Jenkins, F. J. (1995). Use of deoxyinosine in PCR to improve amplification of GC-rich DNA. *BioTechniques*, 19(1), 48–52. PMID: 7669295.

[Udwary *et al.*, 2007] Udwary, D. W., Zeigler, L., Asolkar, R. N., Singan, V., Lapidus, A., Fenical, W., Jensen, P. R., & Moore, B. S. (2007). Genome sequencing reveals complex secondary metabolome in the marine actinomycete *Salinispora tropica*. *Proceedings of the National Academy of Sciences of the United States of America*, 104(25), 10376–10381. PMID: 17563368.

[Unniraman *et al.*, 2001] Unniraman, S., Prakash, R., & Nagaraja, V. (2001). Alternate paradigm for intrinsic transcription termination in eubacteria. *Journal of Biological Chemistry*, 276(45), 41850 –41855.

[Vahedi-Faridi *et al.*, 2010] Vahedi-Faridi, A., Licht, A., Bulut, H., Scheffel, F., Keller, S., Wehmeier, U. F., Saenger, W., & Schneider, E. (2010). Crystal structures of the solute receptor GacH of *Streptomyces glaucescens* in complex with acarbose and an acarbose homolog: Comparison with the acarbose-loaded maltose-binding protein of *Salmonella typhimurium*. *Journal of Molecular Biology*, 397(3), 709–723.

[Van Nieuwerburgh *et al.*, 2011] Van Nieuwerburgh, F., Thompson, R. C., Ledesma, J., Deforce, D., Gaasterland, T., Ordoukhanian, P., & Head, S. R. (2011). Illumina mate-paired DNA sequencing-library preparation using Cre-Lox recombination. *Nucleic Acids Research*.

[Ventura *et al.*, 2007] Ventura, M., Canchaya, C., Tauch, A., Chandra, G., Fitzgerald, G. F., Chater, K. F., & van Sinderen, D. (2007). Genomics of actinobacteria: tracing the evolutionary history of an ancient phylum. *Microbiology and Molecular Biology Reviews: MMBR*, 71(3), 495–548. PMID: 17804669.

[Vera *et al.*, 2008] Vera, J. C., Wheat, C. W., Fescemyer, H. W., Frilander, M. J., Crawford, D. L., Hanski, I., & Marden, J. H. (2008). Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Molecular Ecology*, 17(7), 1636–1647. PMID: 18266620.

[Viswanathan *et al.*, 1999] Viswanathan, V. K., Krcmarik, K., & Cianciotto, N. P. (1999). Template secondary structure promotes polymerase jumping during PCR amplification. *BioTechniques*, 27(3), 508–511. PMID: 10489610.

[VITRESCHAK *et al.*, 2003] VITRESCHAK, A. G., RODIONOV, D. A., MIRONOV, A. A., & GELFAND, M. S. (2003). Regulation of the vitamin B12 metabolism and transport in bacteria by a conserved RNA structural element. *RNA*, 9(9), 1084–1097. PMID: 12923257 PMCID: 1370473.

[Vobis, 1989] Vobis, G. (1989). Actinoplanetes. In H. J. Williams ST, Sharpe ME (Ed.), *Bergy's manual of systematic bacteriology* (pp. 2418–2428).

[Wagner *et al.*, 2009] Wagner, N., Osswald, C., Biener, R., & Schwartz, D. (2009). Comparative analysis of transcriptional activities of heterologous promoters in the rare actinomycete *Actinoplanes friuliensis*. *Journal of Biotechnology*, 142(3-4), 200–204.

[Wang *et al.*, 2011a] Wang, Y., Liu, L., Feng, Z., Liu, Z., & Zheng, Y. (2011a). Optimization of media composition and culture conditions for acarbose production by *Actinoplanes utahensis* ZJB-08196. *World Journal of Microbiology and Biotechnology*.

[Wang *et al.*, 2011b] Wang, Y., Liu, L., Wang, Y., Xue, Y., Zheng, Y., & Shen, Y. (2011b). *Actinoplanes utahensis* ZJB-08196 fed-batch fermentation at elevated osmolality for enhancing acarbose production. *Bioresource Technology*. PMID: 22029955.

[Wang *et al.*, 2009] Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1), 57–63. PMID: 19015660 PMCID: 2949280.

[WATSON & CRICK, 1953] WATSON, J. D. & CRICK, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356), 737–738. PMID: 13054692.

[Watt *et al.*, 2009] Watt, T. F., Vucur, M., Baumgarth, B., Watt, S. A., & Niehaus, K. (2009). Low molecular weight plant extract induces metabolic changes and the secretion of extracellular enzymes, but has a negative effect on the expression of the type-III secretion system in *Xanthomonas campestris* pv. *campestris*. *Journal of Biotechnology*, 140(1-2), 59–67.

[Wawrik *et al.*, 2005] Wawrik, B., Kerkhof, L., Zylstra, G. J., & Kukor, J. J. (2005). Identification of unique type II polyketide synthase genes in soil. *Applied and Environmental Microbiology*, 71(5), 2232–2238. PMID: 15870305 PMCID: 1087561.

[Wehmeier, 2003] Wehmeier, U. F. (2003). The biosynthesis and metabolism of acarbose in *Actinoplanes* sp. SE 50/110: A progress report. *Biocatalysis and Biotransformation*, 21(4-5), 279–284.

[Wehmeier & Piepersberg, 2004] Wehmeier, U. F. & Piepersberg, W. (2004). Biotechnology and molecular biology of the alpha-glucosidase inhibitor acarbose. *Applied Microbiology and Biotechnology*, 63(6), 613–625. PMID: 14669056.

[Weissensteiner & Lanchbury, 1996] Weissensteiner, T. & Lanchbury, J. S. (1996). Strategy for controlling preferential amplification and avoiding false negatives in PCR typing. *BioTechniques*, 21(6), 1102–1108. PMID: 8969839.

[Whiting *et al.*, 2011] Whiting, D. R., Guariguata, L., Weil, C., & Shaw, J. (2011). IDF diabetes atlas: Global estimates of the prevalence of diabetes for 2011 and 2030. *Diabetes Research and Clinical Practice*. PMID: 22079683.

[Winstedt *et al.*, 1998] Winstedt, L., Yoshida, K., Fujita, Y., & von Wachenfeldt, C. (1998). Cytochrome *bd* biosynthesis in *Bacillus subtilis*: Characterization of the *cydABCD* operon. *Journal of Bacteriology*, 180(24), 6571–6580. PMID: 9852001 PMCID: 107760.

[Wurtzel *et al.*, 2010] Wurtzel, O., Sapra, R., Chen, F., Zhu, Y., Simmons, B. A., & Sorek, R. (2010). A single-base resolution map of an archaeal transcriptome. *Genome Research*, 20(1), 133–141. PMID: 19884261.

[Yadav *et al.*, 2003] Yadav, G., Gokhale, R. S., & Mohanty, D. (2003). Computational approach for prediction of domain organization and substrate specificity of modular polyketide synthases. *Journal of Molecular Biology*, 328(2), 335–363.

[Yu *et al.*, 2007] Yu, G., Snyder, E., Boyle, S., Crasta, O., Czar, M., Mane, S., Purkayastha, A., Sobral, B., & Setubal, J. (2007). A versatile computational pipeline for bacterial genome annotation improvement and comparative analysis, with *Brucella* as a use case. *Nucleic Acids Research*, 35(12), 3953 –3962.

[Yuan *et al.*, 2011] Yuan, T., Ren, Y., Meng, K., Feng, Y., Yang, P., Wang, S., Shi, P., Wang, L., Xie, D., & Yao, B. (2011). RNA-Seq of the xylose-fermenting yeast *Scheffersomyces stipitis* cultivated in glucose or xylose. *Applied Microbiology and Biotechnology*, 92(6), 1237–1249. PMID: 22086068.

[Zawilak-Pawlik *et al.*, 2005] Zawilak-Pawlik, A., Kois, A., Majka, J., Jakimowicz, D., Smulczyk-Krawczyszyn, A., Messer, W., & Zakrzewska-Czerwinska, J. (2005). Architecture of bacterial replication initiation complexes: orisomes from four unrelated bacteria. *The Biochemical Journal*, 389(Pt 2), 471–481. PMID: 15790315.

[Zerbino & Birney, 2008] Zerbino, D. R. & Birney, E. (2008). Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Research*, 18(5), 821–829. PMID: 18349386.

[Zhang *et al.*, 2003] Zhang, C., Podeschwa, M., Altenbach, H., Piepersberg, W., & Wehmeier, U. F. (2003). The acarbose-biosynthetic enzyme AcbO from *Actinoplanes* sp. SE 50/110 is a 2-epi-5-epi-valiolone-7-phosphate 2-epimerase. *FEBS Letters*, 540(1-3), 47–52. PMID: 12681481.

[Zhang *et al.*, 2002] Zhang, C., Stratmann, A., Block, O., Brückner, R., Podeschwa, M., Altenbach, H., Wehmeier, U. F., & Piepersberg, W. (2002). Biosynthesis of the C(7)-cyclitol moiety of acarbose in *Actinoplanes* species SE50/110. 7-O-phosphorylation of the initial cyclitol precursor leads to proposal of a new biosynthetic pathway. *The Journal of Biological Chemistry*, 277(25), 22853–22862. PMID: 11937512.

[Zuker, 2003] Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31(13), 3406 –3415.

Appendix A.

A Appendix

A.1. Supplementary figures

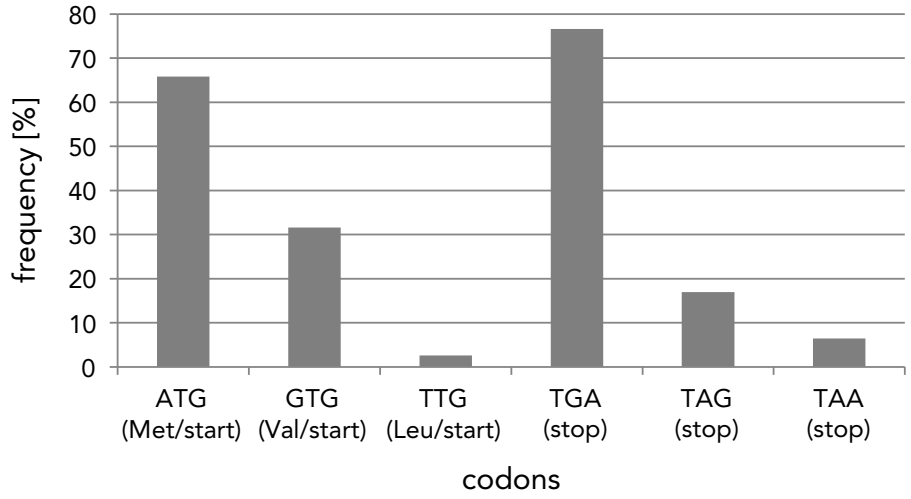

Figure A.1.: Percentage of the three different start and stop codons that are utilized by *Actinoplanes* sp. SE50/110. The depicted data is based on the analysis of all 8.270 CDSs of the organism.

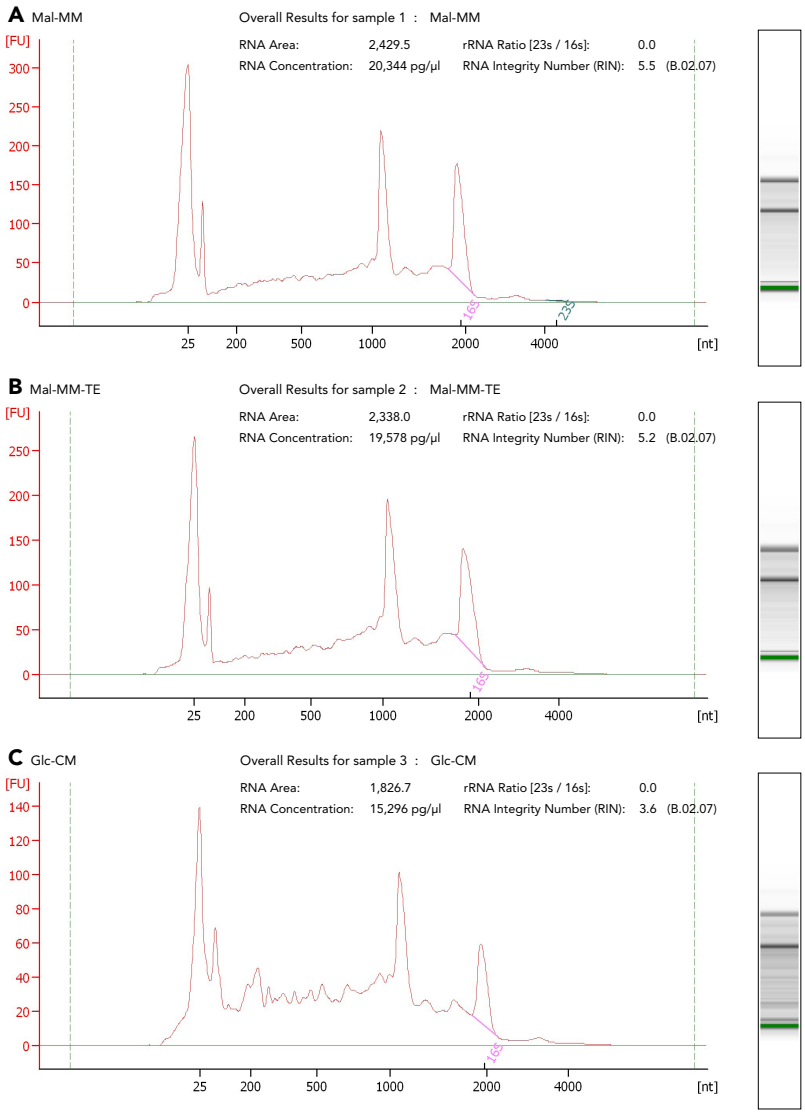

Figure A.2.: Electropherograms for quality assessment of RNA isolation from the three cultivation media **A**) Mal-MM, **B**) Mal-MM-TE, and **C**) Glc-CM for RNA-seq.

A.2. Supplementary tables

Table A.1.: Genes for which the CDS start codon has been corrected.

Gene	TSS position in gene [bp]	Old CDS length [bp]	Old start codon	New CDS length [bp]	New start codon	Trimmed length [bp]
acpl160	5	798	ATG	771	GTG	27
acpl1078	33	1146	GTG	1113	GTG	33
acpl1397	21	933	GTG	912	ATG	21
acpl1860	9	417	ATG	408	ATG	9
acpl1926	9	813	GTG	804	ATG	9
acpl2040	9	414	ATG	405	ATG	9
acpl2065	9	1050	ATG	1041	ATG	9
acpl2292	9	864	ATG	855	ATG	9
acpl2665	24	534	GTG	510	ATG	24
acpl3935	39	462	ATG	423	GTG	39
acpl4578	16	765	ATG	732	GTG	33
acpl4921	21	471	GTG	435	GTG	36
acpl5248	21	288	GTG	243	TTG	45
acpl5968	21	415	ATG	396	ATG	21
acpl6781	39	495	ATG	456	GTG	39
acpl6895	18	774	GTG	756	GTG	18
acpl6975	50	777	ATG	726	TTG	51
acpl7246	39	879	ATG	840	GTG	39
acpl7575	15	729	ATG	711	ATG	18
acpl7780	12	255	ATG	243	ATG	12
acpl7835	9	498	ATG	489	GTG	9
acpl8366	9	876	ATG	867	GTG	9
acpl8063	31	429	ATG	387	GTG	42
acpl8031	30	1293	ATG	1263	GTG	30
acpl8020	3	1419	ATG	1416	GTG	3
acpl7580	9	264	ATG	255	ATG	9
acpl7306	69	1218	ATG	1149	ATG	69
acpl7172	30	1116	TTG	1086	GTG	30
acpl7152	30	1695	ATG	1665	ATG	30
acpl5882	38	414	GTG	195	TTG	219
acpl4957	36	999	ATG	963	GTG	36
acpl4732	18	972	ATG	954	ATG	18
acpl4390	27	861	ATG	834	ATG	27
acpl4348	21	1158	ATG	1137	ATG	21
acpl3416	24	732	GTG	690	ATG	42

Gene	TSS position in gene [bp]	Old CDS length [bp]	Old start codon	New CDS length [bp]	New start codon	Trimmed length [bp]
acpl2248	4	486	ATG	456	GTG	30
acpl1635	9	1179	ATG	1170	GTG	9
acpl1529	33	717	ATG	651	GTG	66
acpl1269	66	948	GTG	714	GTG	234
acpl868	6	402	ATG	396	GTG	6
acpl151	9	495	GTG	486	GTG	9

Table A.2.: Genes for which antisense transcripts have been detected.

Strand	TSS start	Antisense gene	Distance to CDS start	Distance to CDS stop
		TSSs that overlap CDSs of antisense genes		
+	42252	*acpl155*	-220	82
+	42463	*acpl155*	-9	293
+	637469	*acpl663*	-371	633
+	1274365	*acpl1225*	-544	43
+	1658565	*acpl1578*	-463	961
+	1837915	*acpl1736*	-732	80
+	1838071	*acpl1736*	-576	236
+	1987925	*acpl1864*	-1019	117
+	3812700	*acpl3472*	-5	687
+	4392822	*acpl3976*	-206	21
+	4399763	*acpl3985*	-184	22
+	4641805	*acpl4207*	-458	0
+	4983363	*acpl4569*	-165	203
+	6372455	*acpl5846*	-1108	4
+	6693819	*acpl6109*	-70	535
+	6761718	*acpl6152*	-116	1143
+	6775142	*acpl6159*	-539	1170
+	6801261	*acpl6177*	-863	642
+	6841393	*acpl6213*	-1194	620
+	7190136	*acpl6542*	-343	262
+	7253243	*acpl6612*	-419	1116
+	7390081	*acpl6723*	-254	1788
+	7452273	*acpl6775*	-1223	1191
+	8266796	*acpl7473*	-1189	2029
+	8596264	*acpl7783*	-41	870
+	8748225	*acpl7939*	-64	1012
+	8886632	*acpl8068*	-50	1230

Strand	TSS start	Antisense gene	Distance to CDS start	Distance to CDS stop
+	9082279	*acpl8242*	-185	2523
-	8929569	*acpl8104*	-638	39
-	8562262	*acpl7752*	-36	392
-	8414465	*acpl7612*	-56	186
-	7975017	*acpl7217*	-185	9
-	7974964	*acpl7217*	-132	62
-	7974932	*acpl7217*	-100	94
-	7974901	*acpl7217*	-69	125
-	7844523	*acpl7100*	-350	522
-	6347469	*acpl5823*	-54	182
-	5544381	*acpl5089*	-114	275
-	5312093	*acpl4879*	-1774	472
-	5126734	*acpl4715*	-419	522
-	4904552	*acpl4498*	-237	182
-	4729373	*acpl4305*	-668	144
-	4623607	*acpl4187*	-281	537
-	4424076	*acpl4008*	-177	365
-	4376094	*acpl3956*	-627	236
-	4228830	*acpl3809*	-1	259
-	3808039	*acpl3468*	-1292	801
-	3745168	*acpl3414*	-186	335
-	3658510	*acpl3352*	-246	1025
-	3092412	*acpl2864*	-95	219
-	2919770	*acpl2699*	-316	259
-	2412078	*acpl2243*	-986	1179
-	2399789	*acpl2231*	-2353	7
-	2000635	*acpl1876*	-20	435
-	1954232	*acpl1834*	-390	1811
-	1953075	*acpl1833*	-634	571
-	1751505	*acpl1651*	-1291	319
-	1743010	*acpl1642*	-97	832
-	1361616	*acpl1309*	-566	1110
-	1134162	*acpl1097*	-212	1161
-	1121601	*acpl8386*	-128	255
-	1121548	*acpl8386*	-75	308
-	948846	*acpl952*	-858	503
-	869498	*acpl883*	-485	441
-	842316	*acpl860*	-955	304
-	841917	*acpl860*	-556	703
-	784046	*acpl799*	-334	289
-	775449	*acpl781*	-58	250
-	386363	*acpl447*	-223	496

Strand	TSS start	Antisense gene	Distance to CDS start	Distance to CDS stop
-	385911	*acpl446*	-436	232
-	385515	*acpl446*	-40	628
-	169161	*acpl266*	-301	589
-	42163	*acpl154*	-618	2
	TSSs that overlap the promotor region (50 b) of antisense genes			
+	151243	*acpl255*	25	1095
+	955459	*acpl958*	47	1495
+	1316075	*acpl1262*	37	180
+	1532703	*acpl1472*	18	728
+	1646746	*acpl1571*	38	484
+	2563819	*acpl2383*	43	372
+	4038454	*acpl3648*	28	654
+	4438023	*acpl4022*	48	1298
+	7406820	*acpl6737*	9	1463
+	7980143	*acpl7224*	6	863
+	8111274	*acpl7347*	38	1492
-	9215268	*acpl8366*	47	922
-	8868096	*acpl8051*	22	918
-	8863747	*acpl8046*	19	1200
-	8761618	*acpl7957*	12	386
-	8449371	*acpl7646*	34	996
-	8425098	*acpl7624*	35	457
-	7974802	*acpl7217*	30	224
-	6929978	*acpl6300*	4	453
-	6779027	*acpl6162*	42	719
-	6291134	*acpl5750*	44	1036
-	6003779	*acpl5498*	19	1359
-	5204201	*acpl4787*	38	586
-	4836591	*acpl4422*	15	872
-	2695899	*acpl2493*	27	488
-	709741	*acpl720*	43	1887

Table A.3.: Novel non-coding RNAs with unknown function.

Strand	ncRNA gene	Gene start (TSS)	Gene stop	Gene length [bp]
+	*acpl8405*	381882	381974	92
+	*acpl8406*	410489	410752	263
+	*acpl8407*	569122	569427	305
+	*acpl8408*	2717347	2717603	256
+	*acpl8409*	3894052	3894227	175
+	*acpl8410*	8365759	8366023	264
-	*acpl8411*	471460	471324	136
-	*acpl8412*	1956320	1956229	91
-	*acpl8413*	3025969	3025657	312
-	*acpl8414*	5799513	5799217	296
-	*acpl8415*	6922596	6922336	260
-	*acpl8416*	7188946	7188763	183
-	*acpl8417*	7590937	7590339	598
-	*acpl8418*	7918114	7917672	442
-	*acpl8419*	7996667	7996450	217
-	*acpl8420*	8078080	8078023	57
-	*acpl8421*	8697429	8697259	170
-	*acpl8422*	8968167	8967878	289