Computational Methods for the Identification and Characterization of Non-Coding RNAs in Bacteria

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Alexander Herbig

aus Altenkirchen

Tübingen

2014

Tag der mündlichen Qualifikation:	30.01.2015
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatterin:	PD Dr. Kay Nieselt
2. Berichterstatter:	Prof. Dr. Daniel Huson
3. Berichterstatter:	Prof. Dr. Rolf Backofen

Zusammenfassung

Forschungsergebnisse vergangener Jahre konnten zeigen wie komplex die Struktur und Regulation selbst bakterieller Transkriptome sein kann. Auch die wichtige Rolle nicht-kodierender RNAs (ncRNA), die nicht in Proteine translatiert werden, wird dabei immer deutlicher. Diese Moleküle erfüllen in der Zelle verschiedenste Aufgaben wie zum Beispiel die Regulation von Stoffwechselprozessen. Daher ist die Charakterisierung der ncRNA-Gene eines Organismus immer mehr zu einem unverzichtbaren Teil von Systembiologie-Projekten geworden. Hierbei erlauben moderne Hochdurchsatzverfahren im Bereich der DNA- und RNA-Sequenzierung das im hohen Maße detaillierte Studium von Genomen und Transkriptomen. Die daraus resultierenden Daten müssen einer vergleichenden Analyse unterzogen werden, um Variationen des Transkriptoms zwischen verschiedenen Organismen und Umweltbedingungen untersuchen zu können. Hierfür werden effiziente Computerprogramme benötigt, die in der Lage sind genomische und transkriptomische Daten zu kombinieren und entsprechende Analysen automatisiert und reproduzierbar durchzuführen. Zudem müssen diese Ansätze nicht-kodierende Elemente im genomischen Kontext lokalisieren und annotieren können.

In dieser Dissertation präsentiere ich Computerprogramme zur Lösung dieser Aufgaben. So wurde das Programm NOCORNAc entwickelt, welches ncRNAs in bakteriellen Genomen detektiert und diese bezüglich verschiedener Eigenschaften charakterisiert. Dazu gehören zum Beispiel Berechnung von Transkriptionsstart- und endpunkten, Sekundärstruktur und möglicher Interaktionspartner. NOCORNAc wurde im Rahmen einer umfangreichen Transkriptomstudie über das antibiotikaproduzierende Bakterium *Streptomyces coelicolor* verwendet, wodurch die Relevanz von ncRNAs als mögliche Regulatoren gezeigt werden konnte.

Für die komparative Analyse hoch aufgelöster Genom- und Transkriptomdaten multipler Organismen wurde in dieser Dissertation das SuperGenom-Konzept entwickelt, welches bei der vergleichenden Visualisierung multipler Genome Anwendung fand. Zudem diente es als Grundlage für eine neue Methode zur Bestimmung von Transkriptionsstartpunkten in bakteriellen Genomen. Bei der Anwendung auf das für Menschen pathogene Bakterium *Campylobacter jejuni* konnte das Transkriptom dieses Organismus auf globaler Ebene charakterisiert werden. Zudem wurden mehrere bislang unbekannte ncRNAs identifiziert, darunter ein zuvor noch uncharakterisierter CRISPR-Lokus. Hierbei handelt es sich um ein adaptives bakterielles Immunsystem.

Das Studium von Pathogenen kann auch von historischem Interesse sein. Das aufstrebende Feld der Paläogenetik befasst sich mit der Rekonstruktion und Analyse von Genomen alter, mitunter längst ausgestorbener Organismen. In dieser Dissertation werden neue Methoden zur automatischen Rekonstruktion und Charakterisierung alter bakterieller Genome eingeführt, welche zur Erforschung der Evolution von *Mycobacterium leprae* verwendet wurden, dem Verursacher von Lepra.

Die Algorithmen und Werkzeuge, welche in dieser Dissertation entwickelt wurden, sowie die Erkenntnisse, die damit gewonnen werden konnten, stellen einen wertvollen Beitrag zum Verständnis bakterieller Genome und Transkriptome dar und werden weiterhin dazu beitragen deren grundlegende evolutionäre Mechanismen zu verstehen.

Abstract

In recent years the complexity even of bacterial transcriptomes became more and more evident. The important role of so-called non-coding RNAs (ncRNA), which do not encode proteins, is increasingly recognized as they fulfill a variety of functions, such as the regulation of cellular processes or catalysis of other molecules. Therefore, the characterization of an organism's ncRNA repertoire has become an essential part of systems biology studies. In this context novel high-throughput technologies in the field of DNA and RNA sequencing allow for the investigation of genomes and transcriptomes in unprecedented detail. These methodologies produce vast amounts of data that have to be analysed comparatively in order to elucidate variations between different organisms or environmental conditions. For these tasks efficient computational methods are needed that integrate genomic and transcriptomic data from multiple data sets in an automated and reproducible manner. In addition, these approaches have to facilitate the genomic localization of ncRNA elements and their detailed annotation e.g., with respect to promoter regions or transcription start sites as well as their functional characterization such as the prediction of their targets of regulation.

In this dissertation I have made a number of contributions that address these challenges. The computer program nocoRNAc was developed, which predicts ncRNAs in bacterial genomes and characterizes them with respect to multiple properties such as transcription start and end points, secondary structure and potential interaction partners. nocoRNAc has been applied in the context of a comprehensive time series expression study of the antibiotics producing bacterium *Streptomyces coelicolor*, which was cultivated under different environmental conditions. During this study the importance of ncRNAs as potential regulators became evident.

For the analysis of high-resolution genomic and transcriptomic data from multiple organisms the SuperGenome concept was developed. The approach was applied in the context of whole-genome alignment visualization and served as the basis for an algorithm for the comparative detection of transcription start sites in bacterial genomes utilizing RNA-seq data. The application to multiple strains of the human pathogen *Campylobacter jejuni* allowed for the global characterization of this organism's transcriptome and led to the detection of several novel ncRNAs, among them a previously uncharacterized CRISPR locus, which represents an adaptive bacterial immune system.

Studying pathogens can also be of historic relevance. The emerging field of paleogenetics focuses on the reconstruction and analysis of genomes of ancient organisms, whose DNA has been extracted from archaeological samples, such as bones. In this dissertation I present computational methods for the reconstruction and characterization of ancient bacterial genomes, which have been applied to study the evolution of *Mycobacterium leprae*, the bacterium causing leprosy.

Overall, the algorithms and tools developed in this dissertation and the insights that have been gained by their application contribute to the understanding of the structure and organization of bacterial genomes and transcriptomes and will help to elucidate the basic mechanisms that drive their evolution.

Acknowledgements

First of all, I am more than grateful to PD Dr. Kay Nieselt for giving me the opportunity to contribute to so many interesting and exciting projects. I am thankful for the outstanding support I received from her as a supervisor, for encouraging me whenever necessary and giving me a lot of freedom for creative ideas. Not least I would like to thank her for numerous comments and discussions that helped me to improve the quality of my thesis. Furthermore, I want to express my appreciation to Prof. Daniel Huson for being the co-supervisor of my dissertation.

I am owing gratitude to the whole group of Integrative Transcriptomics for support and inspiring discussions on countless occasions, in particular Dr. Florian Battke, Dr. Stephan Symons, Aydın Can Polatkan and Günter Jäger. Special thanks go to Dr. Florian Battke for being a great colleague, for all the joint work in multiple projects and manuscripts, but also for adventurous conference experiences and enlightening discussions about science and beyond.

I am grateful to Sabine Gebert for always being extremely helpful in the organization and planning of uncountable events and activities.

Overall, I would like to thank all my colleagues in the Center for Bioinformatics for an outstanding working environment and for insightful conversations. In addition, I gratefully acknowledge all those bright scientists with whom I had the opportunity to discuss my research at conferences or other occasions. In particular, I am thankful to Prof. Rolf Backofen, Dr. Andreas Richter, Dr. Steffen Heyne, Dr. Fabrizio Costa, Prof. Peter Stadler, Prof. Robert Giegerich, Stefan Janssen, Dr. Aaron Darling and Prof. Peter Wills.

I am very glad that I had the help of very skilled students during multiple projects and especially for exploring and extending the functionalities of the software presented in this dissertation. In particular, I would like to thank Stefan Raue, Daniel Lehle, Mark Rurik, Andreas Friedrich, Linus Backert, André Hennig and Sven Fillinger.

Furthermore, I am thankful to all my collaborators in the SysMO Stream consortium, especially Prof. Wolfgang Wohlleben, Dr. Yvonne Mast, Dr. Eva Waldvogel and Merle Nentwich (University of Tübingen) as well as Dr. Michael Bonin, Dr. Michael Walter (Microarray Facility Tübingen), Prof. Trond Ellingsen (SINTEF, Norway) and Dr. Alexander Wenzel (Norwegian University of Science and Technology).

For the work on GenomeRing I would like to thank Dr. Florian Battke and Günter Jäger, who were concerned with the implementation of the visualizer.

I am very grateful to Dr. Cynthia Sharma, Gaurav Dugar and Dr. Konrad Förstner for the excellent collaboration during the work on *Campylobacter jejuni* and for numerous discussions and suggestions that helped to improve the software developed during this project. Additionally, I would like to thank Prof. Gisela Storz, Maureen Thomason and Thorsten Bischler in the context of the related work on *E. coli.*

For the interesting work on *Staphylococcus aureus* I am thankful to Prof. Friedrich Götz, Dr. Ralph Rosenstein, Dr. Ralph Bertram, Dr. Sabrina Lechner, Christopher

Schuster, Prof. Christiane Wolz, Dr. Wiebke Schröder, Dr. Gabriella Marincola and Dr. Jörg Bernhardt.

I gratefully acknowledge all my co-workers in the medieval leprosy consortium. Most notably, I am grateful to Prof. Johannes Krause, Dr. Dr. Verena Schünemann, Dr. Kirsten Bos, Dr. Ben Krause-Kyora, Prof. Steward Cole, Dr. Pushpendra Singh and Dr. Andrej Banjak. In particular, I am very thankful to Prof. Johannes Krause for the opportunity to be introduced into this extremely exciting field of research.

Last but not least I would like to thank my family and friends for all their support. Most notably, I am extremely grateful to my parents, Rolf and Rosemarie Herbig, for all their love and encouragement.

Contents

List of Figures

List of Tables

1. Introduction

The complex structure of an organism's transcriptome is increasingly recognized as it not only consists of protein-coding transcripts but also of transcripts that are not translated into proteins. For their lacking coding function they are called non-coding RNAs (ncRNAs). Some classes of non-coding RNAs have been known for quite a long time. For example, ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) are well-known housekeeping RNAs. Until today a plethora of further non-coding RNAs have been identified encoded in the genomes of prokaryotes and eukaryotes, which fulfill various functions (see [169, 31] for reviews), such as the regulation of transcription or translation by binding to a target RNA or protein. In addition, they are often involved in catalytic functions, for example in the context of RNA processing.

Considering the wide range of biological mechanisms in which ncRNAs are involved they cannot be neglected in systems biology studies. However, for different reasons they are much harder to identify in genomic sequences than protein-coding genes. As they are not encoding protein sequences they lack an open reading frame. Furthermore, their sequence conservation among species is often significantly lower than it is the case for protein-coding genes. For this reason, computational methods that have been developed for the detection of ncRNAs in genomic sequences usually make use of comparative approaches that – in addition to the sequence – take secondary structure information into account. These approaches are based on the assumption that the secondary structure of ncRNAs is relevant for their function. This makes sense in the context of many possible mechanisms that are employed by RNAs. For catalytic functions it is clear that a specific structure is necessary and regulatory functions that are performed by RNA-RNA interactions potentially also require a certain structure as the interaction site has to be accessible and must not be involved in strong intramolecular interactions. Thus, the detection of conserved secondary structures is a valuable method for the prediction of ncRNAs.

However, not only ncRNA transcripts but also many *cis*-regulatory elements, such as riboswitches, exhibit a conserved secondary structure. Therefore, a conserved secondary structure alone is not a sufficient criterion for the identification of ncRNAs that are transcribed independently of protein-coding messenger RNAs (mRNAs). For this the additional consideration of other criteria is necessary, i.e., the detection of transcriptional signals. Any transcript has a transcription start site or promoter region in general and a transcription terminator signal. Therefore, tools for ncRNA transcript prediction have to include methods for the detection of these features. This dissertation focuses on the characterization of ncRNAs in prokaryotes, where the 3' end of the transcript is often determined by the identification of Rho-independent transcription terminator signals, which are characterized as sequence motifs that lead to the formation of a small stem-loop structure that causes the RNA polymerase to stop transcription. The detection of the ncRNA transcript's 5' start is more challenging. Many approaches are based on the sequence-based identification of sigma factor binding sites such as the program `SIPHT`, for example [93]. However, these sequence motifs are often short and highly degenerated. Furthermore, for many transcription factors the binding site is still uncharacterized. Thus, a more general approach for the detection of the transcription start is desirable.

In this dissertation the computer program NOCORNAC is presented, which predicts and characterizes ncRNA transcripts in bacterial genomes. For the prediction it incorporates methods for the detection of transcriptional signals and the identification of conserved secondary structures. In order to detect promoter regions NOCORNAC does not have to rely on defined transcription factor binding site motifs, but it utilizes an approach for the localization of regions in the genome were the separation and unwinding of the DNA double helix is favorable. This approach is generally applicable to bacterial genomes as it is independent of any information about transcription factors. The 3' end of potential ncRNA transcripts is determined by `TransTermHP` [82], a very fast and accurate method for the detection of Rho-independent transcription terminators. In order to identify conserved secondary structures a pipeline is integrated in NOCORNAC that automatically retrieves homologous sequences for ncRNAs candidates from a sequence database and selects sequences with an optimal evolutionary distance. The candidate ncRNA and the selected sequences are aligned and examined for conserved secondary structures using `RNAz` [61].

A further characterization of predicted ncRNA transcripts is necessary in order to increase the reliability of the prediction and to generate hypotheses about their potential function. For this NOCORNAC incorporates methods for the identification of members of characterized ncRNA families among the predicted elements and for the prediction of RNA-RNA interactions with protein-coding genes.

NOCORNAC has been applied to the prediction of ncRNA transcripts in the genome of the antibiotics producing soil bacterium *Streptomyces coelicolor*. In the SysMO STREAM project this model organism was subject to a comprehensive systems biology study, which assessed the dynamics of the transcriptome, proteome and metabolome of *S. coelicolor* undergoing the metabolic switch from primary to secondary metabolism. For this the wild type and various mutant strains were grown under several nutrient limited conditions. For the transcriptomic studies time series data of unprecedented detail have been produced using a custom design microarray that in addition to protein-coding genes allowed for the expression profiling of ncRNA transcripts predicted by NOCORNAC.

These transcriptome data were not only used to evaluate the predictions but also to comparatively analyse the expression patterns of protein-coding genes and their predicted asRNAs. In addition, putative novel ncRNAs showing a reaction to nutrient depletion were identified. Until then no transcriptome study in *S. coelicolor* had addressed the expression of ncRNAs in such a systematic manner.

Furthermore, RNA-RNA interaction predictions between the predicted ncRNAs and coding genes resulted in a complex network of potential interactions that could be further characterized using NOCORNAC's functionalities for the statistical assessment of RNA-RNA interaction predictions. With this it was possible to identify two predicted ncRNA transcripts that are potentially involved in the global regulation of antibiotics production. These results emphasize the importance of the consideration of non-protein-coding elements in the context of any systems biology study.

Nowadays, deep sequencing technologies (RNA-seq) allow for a much more precise characterization of an organism's transcriptome including the identification of novel protein-coding or non-coding transcripts. The increasing number of available datasets requires novel methods for their comparative analyses. As the data are in single-nucleotide resolution, these methods also have to operate at this level. In addition, comparative analyses have to be conducted for different datasets relating to the same organism, but also between datasets from different organisms. This comparison, however, is challenging as genomes of different organisms usually differ significantly due to insertions, deletions or genomic rearrangements. Thus, the datasets that are subject to the comparative analysis relate to different coordinate systems.

As a solution to this problem the SuperGenome concept is presented in this dissertation. The data structure of the SuperGenome is constructed on the basis of a multiple whole-genome alignment and provides a common coordinate system for multiple genomes, which allows for comparative analyses of genomic or transcriptomic data that relate to different genomic coordinate systems.

The SuperGenome can serve as the basis for various applications. As the underlying data structure not only allows for the representation of insertions and deletions but also models genomic rearrangements, the SuperGenome can be applied to the comparative analysis of genomic architectures. In this context GenomeRing is described in this dissertation. GenomeRing is a circular visualization method for the identification of architectural differences and similarities between genomes. The SuperGenome forms the algorithmic foundation of GenomeRing, thus allowing for the projection of genomic annotations into the SuperGenome coordinate system, which can also be visualized in GenomeRing. Furthermore, the integration of GenomeRing with the transcriptome analysis software Mayday [14] makes detailed analyses of single loci within one framework possible that can also integrate experimental data from various sources.

The coordinate mapping that is produced by the SuperGenome approach has single-nucleotide resolution. Therefore, any comparative data analysis that is performed on the basis of the SuperGenome can be conducted in single-nucleotide resolution. Thus, the SuperGenome approach is ideal for the analysis of RNA-seq data. The assessment of RNA-seq data leads to a detailed characterization of an organism's transcriptome as transcript boundaries can be precisely defined.

One important information that can be gained from RNA-seq data is the exact location of transcription start sites (TSS) of protein-coding or non-protein-coding genes. In addition, it is possible to identify novel transcripts. Previously the annotation of transcription start sites on the basis of RNA-seq data has been performed manually in most cases [142]. This approach, however, is extremely time consuming and has a low reproducibility. Furthermore, it becomes almost infeasible for the comparative analysis of multiple data sets involving different organisms.

Thus, computational methods are required that allow for an automated reproducible and comparative annotation of TSS across multiple data sets from different organisms. In this thesis an algorithm for the computational detection of TSS was developed, which has been combined with the SuperGenome approach to allow for its application to data that is related to different genomes. By the integration of the SuperGenome, TSS annotated to different genomes can be associated to each other even if the genomes differ in terms of insertions, deletions or genomic rearrangements.

The TSS detection algorithm together with the SuperGenome approach have been combined with a user-friendly graphical user interface (GUI). This software tool, called TSSpredator, allows the user to set all relevant parameters, which makes the procedure accessible for scientists without large expertise in computer science.

TSSpredator has been applied to the comparative detection and analysis of TSS in four *Campylobacter jejuni* strains. The genomes of these strains differ significantly by large insertions. The SuperGenome approach, however, allows for the consistent assignment of coordinates within the SuperGenome coordinate system, thus making a comparative analysis of the TSS possible despite the variation of the genomic architectures.

The analysis of the global TSS maps of the four strains allowed for a global characterization of promoter regions and the identification of SNP-dependent promoter usage that differs between the strains. Furthermore, several candidates for novel non-coding RNAs were discovered including a new CRISPR locus.

This shows how comparative analyses can integrate genomic information with transcriptomic data to provide new insights into regulatory mechanisms that potentially differ between different bacterial strains and that might even influence phenotypic variations e.g., in the context of pathogenicity.

Transcriptomic information, however, is not available in all fields of research. For example, in the emerging field of paleogenetics, which deals with the analysis of ancient DNA [171], studies are limited to the investigation of genomes as RNA is usually not preserved. Thus, comparative methods have to rely on genomic information alone to generate hypotheses about phenotypic differences between ancient and modern organisms and about the evolution of the investigated species.

The analysis of DNA that has been isolated from old samples is significantly more challenging than for modern isolates as ancient DNA (aDNA) is degenerated, which means that the DNA fragments are very short and the molecules are damaged in a way that leads to wrong base calls during sequencing. Furthermore, ancient samples have to be regarded as metagenomic samples in almost all cases, because the DNA of various modern organisms but potentially also ancient organisms is contained in the sample in addition to the DNA of the target organism. Though DNA capture techniques are applied for the enrichment of the target organism's DNA, the DNA of other organisms is still contained, which has to be considered during the computational analysis of the data.

Tools for the specific preprocessing of aDNA sequencing data have been published earlier [83]. In addition, software has been developed that considers the properties of aDNA sequencing reads during mapping [58]. However, these methods only cover parts of the required analysis steps and are in many cases not efficient enough for the application to full bacterial genomes.

Therefore, a computational pipeline for the efficient comparative genomic analysis of ancient and modern bacterial strains was developed in this dissertation. The steps covered by this pipeline include preprocessing of the DNA sequencing reads, mapping to a reference genome, SNP calling based on the mapping, filtering of called SNPs and a SNP effect analysis to estimate the influence on protein-coding genes. This can lead to hypotheses about phenotypic differences in comparison to modern strains of the same species. Furthermore, draft genome sequences are generated on the basis of the detected SNPs and the draft genomes of sequenced ancient and modern strains are aligned with published reference strains to allow for a comparative SNP analysis. This information can then be used for phylogenetic and dating analyses.

This dissertation describes the application of this pipeline to the comparative genomic analysis of ancient and modern strains of *Mycobacterium leprae*, the bacterial pathogen causing leprosy. In this study the genomes of medieval and modern *M. leprae* strains were sequenced and comparatively analysed in order to investigate evolutionary relationships between the strains and to elucidate the origin and history of this bacterium in the context of its role as a human pathogen. By using the pipeline which is introduced here these analyses could be conducted in an efficient and reproducible manner.

Overall, this thesis presents multiple algorithms and tools for the characterization of non-coding as well as protein-coding transcripts in bacterial organisms. These methods together with the knowledge that could be gained by their application form a valuable contribution to the fields of genomics as well as transcriptomics and will assist researchers in elucidating the architectures of genomes and transcriptomes of bacteria while at the same time forming the basis for future development.

This dissertation is structured as follows: A biological background as well as an overview of the computational methods relevant for this thesis are presented in chapter two. Chapter three describes the non-coding RNA prediction and characterization software NOCORNAC, which was published in 2011 [65]. NOCORNAC has been applied to the identification of non-coding RNAs in the genome of *Streptomyces coelicolor*, which have been verified and further characterized by a comprehensive time series transcriptomic study (chapter four), that is presented in multiple publications [109, 13, 65, 164]. The SuperGenome approach and its application to the visualization of genomic architectural variation is described in chapter five. It won the *Most Creative Algorithm Award* of the *Illumina iDEA Challenge 2011* and was presented at the ISMB 2012 [64]. Chapter six describes a novel algorithm for transcription start site prediction on RNA sequencing data, which integrates the SuperGenome approach for a comparative analysis of multiple genomes. This algorithm was published in 2013 together with its application to the comparative detection and characterization of transcription start sites in four *Campylobacter jejuni* strains [45] (chapter seven). Chapter eight describes a computational paleogenetics pipeline which was applied to the analysis of medieval and modern strains of *Mycobacterium leprae* [141]. Chapter nine provides a discussion of the work presented in this dissertation.

2. Background

2.1. Non-coding RNAs and their regulatory function in bacteria

In bacteria non-coding RNAs (ncRNA) fulfill a plethora of functions by various mechanisms [169, 31].

The most well known ncRNAs are the housekeeping RNAs. These are, for example, the transfer RNAs (tRNA), which guide single amino acids to the translational machinery and represent the link between the codons and the respective amino acids. They are, therefore, an essential part of the genetic code itself. Another class of housekeeping RNAs is part of the aforementioned translational machinery. These are the ribosomal RNAs (rRNA), which form the ribosome together with ribosomal proteins. One further example of a housekeeping RNA is the transfer-messenger RNA (tmRNA). The function of this ncRNA is to release stalled ribosomes from messenger RNA (mRNA) molecules that became non-functional due to degradation, for example. Other housekeeping RNAs are the RNA component of the signal recognition particle or RNase P, which is involved in tRNA processing.

These housekeeping RNAs can be found in all bacteria. Another diverse group of ncRNAs, which are very often not conserved over a wide range of species, is the group of regulatory RNAs. In most cases they regulate the expression of their target genes by base-pairing interactions between the ncRNA transcript and the mRNA of the target gene. This interaction can influence the translation of the mRNA or the stability of the target transcript. However, there are also ncRNAs that bind to proteins and directly regulate their activity.

Regulatory RNAs can be encoded at the same locus as their target mRNA overlapping the coding region or at least the untranslated region of the target transcript. Therefore, large parts of the two transcripts have a complementary sequence. These ncRNAs are called *cis*-antisense RNAs (asRNAs). In contrast, *trans*-antisense RNAs are encoded at a different locus, but their sequences are still partially complementary to the respective sequences of their target transcripts.

In addition to these regulatory ncRNA transcripts there is another group of *cis*-regulatory elements, which are not transcribed on their own, but which are part of mRNAs. They can act as signals but also as sensors influencing the secondary structure of the mRNA molecule they are part of. These so-called riboswitches can thereby activate or inactivate the translation of the respective gene.

The different groups of regulatory RNAs and their mechanisms of regulation are described in more detail in the following sections.

2.1.1. Riboswitches

Riboswitches and other *cis*-regulatory elements are part of the 5' untranslated region (UTR) of mRNAs regulating their transcription or translation [169, 63]. Thus, they are not transcribed on their own but as part of another transcript. They are still considered as functional RNAs, because they have a specific secondary structure, which is essential for their regulatory function. As these elements are located in the 5' UTR, they are also called 'leaders'. One example is the 'T-box leader', which is part of the mRNA of tRNA synthetases. This element binds the respective uncharged tRNA, which leads to a repression of the gene and thereby to a downregulation of tRNA production. Another type of leader elements are the 'RNA thermometers', which regulate the expression of the downstream gene according to the temperature.

A group of leader sequences that is able to bind metabolites and perform the regulation of genes involved in the uptake and processing of these molecules are called 'riboswitches'. A riboswitch consists of an 'aptamer', where the ligand is bound, and the 'expression platform', which changes its structural conformation in the presence of the ligand and thereby regulates the expression of the gene which is located downstream of the riboswitch. This usually involves the formation of hairpin structures that either act as transcription terminators or antiterminators, or they lead to the occlusion of the ribosome binding site or make it accessible. Thus, riboswitches can regulate transcription and translation as activators or repressors. However, the inactivation of gene expression is found much more often.

Interestingly, it is even possible that the same type of riboswitch has different regulatory mechanisms depending on the species [169, 111]. For example, the cobalamin riboswitch binding coenzyme B_{12} acts as a transcription terminator in Gram-positive bacteria while regulating translation in Gram-negative bacteria.

It is also possible that transcripts carry multiple riboswitches. In general it turned out that regulation by *cis*-regulatory elements occurs more often in Gram-positive bacteria. In *Bacillus subtilis*, for example, about 2% of the genes are regulated by riboswitches.

2.1.2. cis-encoded antisense RNAs

Regulatory ncRNAs acting by base-pairing with their target transcript are either *cis*-encoded or *trans*-encoded. In the case of *cis*-encoded antisense RNAs (asRNA) the asRNA is encoded on the opposite strand of the regulated gene, but at the same locus, so that the two transcripts are overlapping and their sequences in the overlap region are complementary to each other [169, 26]. The size of the overlap, which is complementary, can be quite large, often more than 75 nucleotides. Although the two transcripts are *cis*-encoded, they are *trans*-acting in the bacterial cell, i.e., they are transcribed independently from each other and get in contact by diffusion. Due to their complementary sequence they form a stable duplex. In most cases this leads to the occlusion of the ribosome biding site of the target mRNA. There are also mechanisms known where the duplex formation with the asRNA leads to a faster degradation of the mRNA [169].

The asRNA can, however, also influence the transcription of the target gene. The continuous transcription of the asRNA can lead to a downregulation of the transcription of the target mRNA on the opposite strand [54]. In an operon, where the asRNA is encoded around the 3' end of an open reading frame (ORF), the asRNA can bind to the transcript of the operon before transcription is finished, which leads to a pre-mature transcription termination and, therefore, the downregulation of the genes that are encoded downstream of the asRNA's binding site.

A specific group of genes regulated by asRNAs are type I toxin-antitoxin systems [169, 161]. These systems consist of a protein-coding transcript that encodes a toxic protein and an asRNA inhibiting the translation and promoting the degradation of the toxin-encoding mRNA. Thus, the asRNA is the antitoxin of that system. These systems are usually found on plasmids and ensure that the plasmid is preserved in the population. If a daughter cell of the bacterium does not contain the plasmid the less stable antitoxin asRNA will be degraded and no further transcripts can be produced since the plasmid was lost. The more stable mRNA of the toxin, however, is no longer inhibited. Thus the toxic protein is produced, which kills the bacterial cell that has lost the plasmid. Interestingly, toxin-antitoxin systems have also been found on bacterial chromosomes. The role of these chromosomally encoded systems is still unclear, but it has been shown that many toxins might have regulatory functions when they are present in lower concentrations, e.g., by causing a slower cell growth [80, 159].

2.1.3. trans-encoded ncRNAs

If the regulatory ncRNA is transcribed from a different locus than the mRNA of its target, it is called a *trans*-encoded regulatory RNA. The complementary region between these ncRNAs and their target transcripts is often significantly shorter than for *cis*-asRNAs and involves only 10 to 25 nucleotides [169]. The regulatory mechanisms, however, are similar. Like the *cis*-asRNAs the *trans*-ncRNAs bind to their target and occlude the ribosome binding site, which inhibits translation. In addition, the degradation of the mRNA is promoted in most cases. Nevertheless, in some cases the binding of the ncRNA stabilizes the mRNA or it can even activate its translation. This happens if the native structure of the mRNA prevents the ribosome from attaching and the binding of the ncRNA changes the structural conformation and makes the ribosome binding site accessible. These ncRNAs are also known as 'anti-antisense RNAs'.

Unlike *cis*-asRNAs *trans*-encoded regulatory ncRNAs often have multiple targets. Thus, they act like transcription factors regulating a whole group of genes but doing this on a post-transcriptional level. In addition, they are often expressed at very specific growth conditions whereas many *cis*-asRNAs are transcribed constitutively.

For the RNA-RNA interaction between the ncRNA and the mRNA the RNA chaperone Hfq is required in many cases. It binds the ncRNA stabilizing it and also catalyzes the RNA-RNA interaction by changing the conformation of the RNAs to make the binding sites accessible. Hfq might also be involved in promoting the degradation of the mRNA once the ncRNA is bound. It has been shown, however, that *trans*-encoded ncRNAs do not necessarily become ineffective in Hfq mutant strains. Furthermore, there are bacteria without a Hfq homolog that express functional regulatory ncRNAs.

2.1.4. ncRNAs regulating protein activity

In addition to regulation via RNA-RNA interaction there are ncRNAs that bind to proteins and thereby regulate the protein activity. CsrB and CsrC, for example, are two ncRNAs that modulate the activity of the protein CsrA in *E. coli.* CsrA binds to mRNAs that contain a GGA motif in their 5' UTR, which influences the stability and translation of these mRNAs. The CsrB and CsrC ncRNA transcripts contain multiple instances of the same motif, by which they are able to bind CsrA proteins. Thereby, the CsrA proteins are blocked and are not able to bind to their target mRNAs. Thus, the expression of CsrB and CsrC leads to a global downregulation of CsrA activity. The two ncRNAs are in turn regulated by the CsrD protein, which binds to them and promotes their degradation by RNase E.

Another example for a regulation of protein activity by an ncRNA is the 6S RNA in *E. coli.* The 6S RNA mimics an open promoter and binds and thereby sequesters RNA polymerases that are in complex with the σ^{70} transcription factor. However, it does not bind RNA polymerases which are in complex with the σ^{S} transcription factor. By this mechanism the expression of some genes with σ^{70} promoters is downregulated while the expression of some genes with σ^{S} promoters is upregulated during stationary phase when 6S RNA is highly abundant.

2.1.5. CRISPR RNAs

A very specialized type of RNAs are CRISPR RNAs. The CRISPR (clustered regularly interspaced short palindromic repeats) system can be regarded as a kind of immune system found in bacteria and archaea [169, 79, 152]. It is able to silence foreign DNA that has entered into the cell such as from phages or plasmids. CRISPR loci are highly dynamic and can adapt to phages the bacterium has been infected with and incorporate elements to gain immunity against these invaders.

A CRISPR locus consists of several short repeat units, which are interspersed with short spacer units. The length of the repeats is 21 to 47 nucleotides while the spacers are between 20 and 72 nucleotides in size. The number of repeat-spacer units is usually between 20 and 30 but can be greater than 200 in some cases. While the sequence of the repeats is highly conserved, the spacer sequences are extremely diverse. The array of repeat-spacer units is preceded by a so-called 'leader sequence', which is conserved and relevant for the functionality of the system. Furthermore, the locus is surrounded by CAS genes, which encode the protein components of the CRISPR system. The spacers, which are complementary to parts of foreign DNA, are actually the carriers of immunity information.

When foreign DNA enters the cell, e.g., in the form of a plasmid or during a phage infection, CAS proteins incorporate small pieces of this DNA (protospacer) into the CRISPR locus, where they form a new spacer. The incorporation takes place between the leader sequence and the first repeat-spacer unit. Thus, the CRISPR array grows at only one side whereas elements on the other side originate from older infections.

The whole CRISPR array is constitutively transcribed. One long transcript is produced, that contains all repeat-spacer units. A complex of CAS proteins, however, processes this transcript and cleaves it into smaller RNAs, where each contains one single repeat-spacer unit. These are the CRISPR RNAs (crRNA). The crRNAs are bound by a complex of CAS proteins. When foreign DNA enters the cell, crRNAs that show complementarity to the DNA bind to it and the CAS proteins, which build a complex with the crRNA, promote the degradation of the invading DNA.

2.1.6. Multifunctional RNAs

Although most RNAs probably fulfill a single specific function such as either protein-coding or as a non-coding regulator, there are several examples of RNAs known that facilitate multiple functions. RNA III in *S. aureus*, for example, regulates virulence factors by RNA-RNA interaction with their mRNAs but at the same time encodes a small protein of 26 amino acids length. Another example is the SgrS RNA in *E. coli*, which downregulates the expression of PtsG by RNA-RNA interaction with its mRNA. PtsG is a sugar-phosphate transporter. In addition, SgrS encodes the 43 amino acid protein SgrT, which also inhibits the PtsG transporter. Thus, the SgrS RNA acts as a downregulator of glucose uptake by two different regulatory mechanisms targeting the same gene.

2.2. Computational prediction and characterization of ncRNAs

2.2.1. Overview

Computational approaches to the detection of ncRNAs can be divided in three different groups depending on the type of elements that are to be detected [67, 110, 10, 102, 146]. An overview of the most important approaches is given in figure 2.1.

The first group deals with the *de novo* detection of ncRNAs, which aims at the prediction of ncRNA elements in general without any knowledge about certain sequential or structural features of the elements. This makes the *de novo* prediction of ncRNAs the most challenging problem in the field of ncRNA detection. Most algorithms applied to this problem make use of methods for comparative sequence analysis for the detection of sequence motifs or the detection of conserved secondary structures. For the detection of ncRNA transcripts these approaches are complemented by methods for the detection of transcriptional features such as promoter regions and transcription terminators.

The second group of algorithms deals with the detection of elements that belong to a certain RNA family. If the family is already defined by a group of sequences, other members of the same family can be detected by sequence similarity, for example. If, however, the pairwise sequence similarity within the family is too low, secondary structure information that is available for the members of the RNA family can be

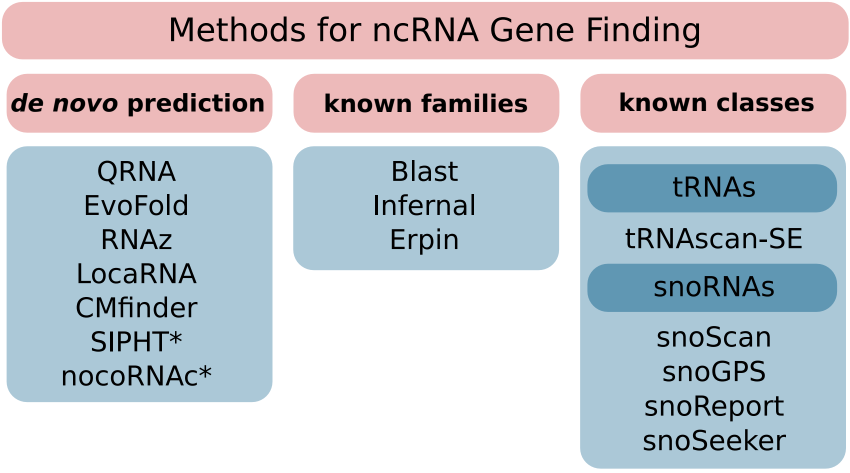


Figure 2.1.: Overview of the three approaches to ncRNA prediction and examples of relevant programs. 1. *De novo* prediction, for which a subgroup of programs (*) is available that specifically predict ncRNA transcripts. 2. Searching for members of a specific family. 3. Searching for members of a specific class. These programs are usually specialized for single classes of ncRNAs.

used to train a probabilistic model which integrates sequence and structure information. This model can then be applied to a set of target sequences or whole genomes to detect further members of the same RNA family.

The third group of algorithms were developed for the detection of ncRNAs that belong to a specific RNA class. Members of an ncRNA class show very low sequence conservation while sharing similar structural elements and facilitating a similar function. Because these structural patterns are very specific for an individual RNA class, many algorithms are designed for the detection of only one specific class. An example for an RNA class are microRNAs (miRNAs). The sequence conservation among different miRNAs is very low, but due to their distinct structural features several programs could be developed that specifically detect miRNAs.

Most of the methods in the field of ncRNA detection can be applied to long target sequences, e.g., complete genomes, to detect ncRNAs *de novo* or as members of known families or classes. However, these methods can also be used for RNA classification. Here, several methods are applied to a set of RNA sequences to decide if they belong to a certain family or class or if they might represent a novel class of RNAs.

2.2.2. De novo prediction of non-coding RNAs

Methods based on comparative sequence analysis

For the *de novo* detection of ncRNAs, which is not limited to a specific family or class of RNAs, approaches have been developed that are based on comparative sequence analysis. These methods search for conserved secondary structures and sequence patterns in a set of homologous sequences. Depending on the applied program these sequences are either aligned as part of the prediction process or they have to be aligned prior to the application of the program. This can be accomplished by using `ClustalW` [154], for example. However, as a first step the set of homologous sequences has to be defined. If the detection of ncRNAs is applied to a complete genome, a multiple whole-genome alignment using the genomes of organisms closely related to the target organism can be generated by using whole-genome aligners such as Mauve [38]. If the search is restricted to specific regions or elements, such as intergenic regions or experimentally detected transcripts, homologous sequences are gathered from large sequence databases, e.g., by using the basic local alignment search tool (BLAST [6]).

The program `RNAz` makes use of such a comparative approach for the *de novo* detection of ncRNAs [61]. It takes a multiple alignment of homologous sequences as input and predicts if the aligned sequences contain a structurally conserved ncRNA. For long sequences, like a whole-genome alignment, for example, `RNAz` is applied using a sliding window approach. The program makes use of a support vector machine (SVM) to classify the input. The SVM considers various properties of the input sequences. The two most important are the z-score and the structure conservation index (SCI). The z-score is defined as follows:

$$z = \frac{m - \mu}{\sigma}, \tag{2.1}$$

where m is the average minimum free energy (MFE) of the secondary structures predicted for the sequences in the alignment and μ and σ are mean and standard deviation of the MFE values of the structures of random sequences, which have a similar length and dinucleotide composition as the input sequences.

The z-score is a measure of the stability of the input sequences when compared to the expected stability of random sequences with similar properties. The assumption here is that the secondary structure of a functional RNA is significantly more stable than the structure of sequences that do not contain a functional structure. Thus, large negative z-scores indicate that the input sequences are significantly more stable than expected by chance.

The second assumption is that the secondary structure of functional RNAs is not only more stable but also significantly more conserved than that of other sequences. The degree of structure conservation among the aligned input sequences is measured by the structure conservation index (SCI):

$$SCI = \frac{E_A}{\overline{E}}, \tag{2.2}$$

where E_A is the MFE of the consensus structure of the aligned sequences, which is calculated by `RNAalifold` [19], and $\overline{E}$ is the average MFE of the predicted secondary structures of the single sequences.

If the secondary structures of the single sequences are quite similar, their MFE values will be similar to the MFE of the consensus structure and thus, the SCI will be close to 1. If, however, the individual structures are very dissimilar, their MFEs will be significantly lower than the MFE of the consensus structure, which leads to a SCI close to 0.

The z-score, the SCI and other properties of the alignment, such as the length and the mean pairwise sequence identity, are used as input for the SVM classifier. As a result `RNAz` provides for each input alignment the probability that this alignment contains a structured RNA.

A different comparative approach is implemented in the program `EvoFold` [119]. It utilizes phylogenetic stochastic context-free grammars (phylo-SCFG) to detect functional RNAs. Like `RNAz` `EvoFold` takes aligned sequences as input. It applies two different models to the input, where one model represents functional RNAs and the other one is a background model. Both models are implemented as phylo-SCFGs. With SCFGs, it is not only possible to model sequence patterns as it is done with hidden Markov models (HMMs), but it is also possible to model secondary structure information. Additionally, a phylogenetic tree modelling the divergence among the input sequences is taken as input, which is used for an individual weighting of nucleotide substitutions of single sequences within the alignment.

Another program that utilizes SCFGs is `QRNA` [129], but it does not make use of any phylogenetic information. In addition, only the model for functional RNAs is based on SCFGs while protein-coding sequence and background are modelled with HMMs. The input for `QRNA` are two aligned sequences, which are assigned to one of the three classes as defined by the three models: ‘functional RNA’, ‘protein-coding’ and ‘other’.

There are also programs, which take unaligned sequences as input and perform simultaneous sequence and structure alignment of the input data. One of these programs is `LocaRNA` [170]. `LocaRNA` calculates conservation profiles for sequence and structure in single-nucleotide resolution. With `LocaRNA` it is therefore possible to determine the exact boundaries of conserved RNA structures within the input sequences.

`CMfinder` is another program that takes unaligned sequences as input [178]. It can be seen as a motif finder for secondary structure motifs. It employs covariance models (CMs), which are based on SCFGs, to describe the structural motifs that have been identified in the input sequences. These CMs can be used to search for the respective motifs in genomic sequences or other sequence data.

Another program that assesses structural conservation and that can be used for ncRNA prediction is `Dynalign` [160].

Methods based on transcriptional feature detection

Methods based on comparative sequence analysis that detect functional RNAs do not distinguish between structural elements, which are part of messenger RNAs, and ncRNAs that are transcribed on their own. *Cis*-regulatory elements, for example, are part of mRNAs, but also Rho-independent transcription terminator signals. To be able to identify mRNA-independent ncRNA transcripts, the prediction of transcriptional features, such as promoter regions and transcription terminators, is combined with the detection of conserved secondary structures.

One program following this approach is SIPHT [93]. It provides a web-based interface for its application to various bacterial genomes. SIPHT predicts ncRNA transcripts only in intergenic regions. Homologous sequences are identified by comparing the intergenic regions of the target organism with other bacteria using BLAST [6]. For candidate regions transcription factor binding sites (TFBS) are detected with position-specific weight matrices (PSWM) and Rho-independent terminator signals are predicted by utilizing the program TransTermHP [82]. For the detection of conserved secondary structures QRNA is integrated in the SIPHT pipeline.

nocoRNAc [65], another program for the detection of ncRNA transcripts, is presented in this thesis (section 3). nocoRNAc is not restricted to intergenic regions, but also able to detect antisense RNAs. For the prediction of promoter regions it makes use of a thermodynamic model for the calculation of DNA duplex stability, which does not have to rely on known TFBS motifs. Furthermore, nocoRNAc integrates TransTermHP [82] for terminator prediction, RNAz [61] for the detection of secondary structure conservation and IntaRNA [30] for the prediction of RNA-RNA interactions.

Other programs taking heterogeneous data such as transcription signals into account are sRNAfinder [155], sRNApredict [92] or sRNAscanner [147].

Further approaches for the *de novo* detection of ncRNAs are based on sequence clustering [157], graph processing [32] or different machine learning approaches [156, 132, 175, 133].

2.2.3. Searching for members of an RNA family

For the detection of RNAs that belong to an individual RNA family a specific model can be trained for that family that can then be employed for searching other family members. A database where models of RNA families are stored is Rfam [29]. Part of Rfam is the Infernal software package, which contains programs for building family models from multiple sequence alignments (CMbuild) or for searching members of a given family (CMsearch). For each family an alignment of family members, a consensus secondary structure and the respective covariance model can be found on Rfam. CMsearch takes a CM and a target sequence as input and searches for instances of the respective RNA family in the target sequence. It lists all matches including coordinates, bit score and an e-value. CMsearch can apply various filtering steps prior to the application of the CM, which reduces the time-consumption. An alignment HMM (HMMer), which models the multiple alignment of known family members, can be used to identify regions in the target sequence that show sufficient sequence similarity, so that only for these candidate regions the full CM is applied. Alternatively, BLAST can be used to identify candidate regions.

A program for RNA family search that is not based on CMs is Erpin [51]. For a given alignment and consensus structure of members of an RNA family as it can be retrieved from Rfam, Erpin constructs position-specific weight matrices (PSWM) for structural elements like stems and loops. The PSWMs are then used to scan a target sequence (e.g., a genome) for instances of this RNA family. Erpin needs a descriptor file of the query RNA family as input, which can, however, be automatically generated from Rfam entries. A manual modification of the descriptor file is also possible, for example to remove structural elements from the descriptor that are not well conserved.

2.2.4. Searching for members of an RNA class

Most programs for the detection of ncRNAs that belong to a certain class are restricted to only one specific class or a few related classes. One class of RNAs that is also regarded as an RNA family is the class of tRNAs. A program for the detection of tRNAs in genomic sequences is tRNAscan-SE [94, 135]. It uses different models for the prediction depending on the target organism (bacteria, eukaryotes, etc.).

Another class of RNAs for whose detection specific methods have been developed is the class of small nucleolar RNAs (snoRNAs), which are involved in the processing of small nuclear RNAs (snRNAs) and ribosomal RNAs (rRNAs). Programs for their prediction rely on the detection of the specific sequence motifs in C/D box snoRNAs and H/ACA box snoRNAs and on the localization of certain structurally conserved elements. Parts of their sequence are complementary to their target RNAs (snRNAs, rRNAs) and the detection of these complementary regions is also part of some programs for snoRNA detection.

The program snoscan [95, 135], for example, considers such target information for the detection of C/D box snoRNAs. A specific prediction of H/ACA box snoRNAs is accomplished by snoGPS [136, 135]. The program SnoReport [68] predicts both, C/D box and H/ACA box snoRNAs, and does not need any target information. A combination of two approaches is snoSeeker [177], which integrates CDseeker for the detection of C/D box snoRNAs and ACAseeker for the prediction of H/ACA box snoRNAs. The program can be applied to genome alignments but also to sequencing data.

3. nocoRNAc: Prediction and characterization of non-coding RNAs

Bacterial non-coding RNAs (ncRNAs) are increasingly recognized as key regulators that are involved in various biological processes (see section 2.1). Therefore, several methods for the computational prediction and characterization of ncRNAs have been developed. An overview is provided in section 2.2. Most of these methods focus either on the detection of ncRNAs that are similar to known RNA families or classes, or they are based on the detection of conserved secondary structures. Few methods incorporate the detection of transcriptional features for the prediction of ncRNA transcripts, which is based on the identification of known transcription factor binding site (TFBS) motifs and Rho-independent terminator signals. In addition, most approaches are limited to intergenic regions and ignore *cis*-antisense RNAs.

In this chapter the program nocoRNAc is presented, which is designed for the genome-wide prediction and characterization of ncRNA transcripts in bacteria. It integrates methods for the detection of conserved secondary structures with the identification of transcriptional features. However, nocoRNAc does not rely on described TFBS but utilizes a more general model (SIDD) for the localization of promoter regions, which is described in section 3.1.2. A first version of nocoRNAc is described in my diploma thesis [66]. It consisted of basic methods for the classification of ncRNA candidates including a first implementation of the SIDD model. In this dissertation the SIDD approach was partially reimplemented for higher efficiency. In addition, methods for the genome-wide detection of structured ncRNA transcripts including an automated structure conservation pipeline were developed. Furthermore, nocoRNAc now provides functionalities for the more detailed characterization of ncRNA candidates such as the prediction of RNA-RNA interactions (section 3.4). All results of an application of nocoRNAc can be assessed in its interactive R environment (section 3.5). An overview of nocoRNAc's workflow is depicted in figure 3.1.

In general nocoRNAc employs three different strategies for the prediction of ncRNA transcripts.

The first strategy is to take regions as input that potentially contain conserved RNA secondary structures, which have been identified with RNAz [61] by the user. For this RNAz is applied to a whole-genome alignment of the target organism's genome with one or more other genomes of related organisms. The resulting regions are further processed by nocoRNAc. Rho-independent terminator signals and SIDD sites that have been predicted by nocoRNAc are assigned to the regions identified by RNAz and combined to predict ncRNA transcripts in the context of these regions (section 3.2). In general nocoRNAc is not limited to process regions

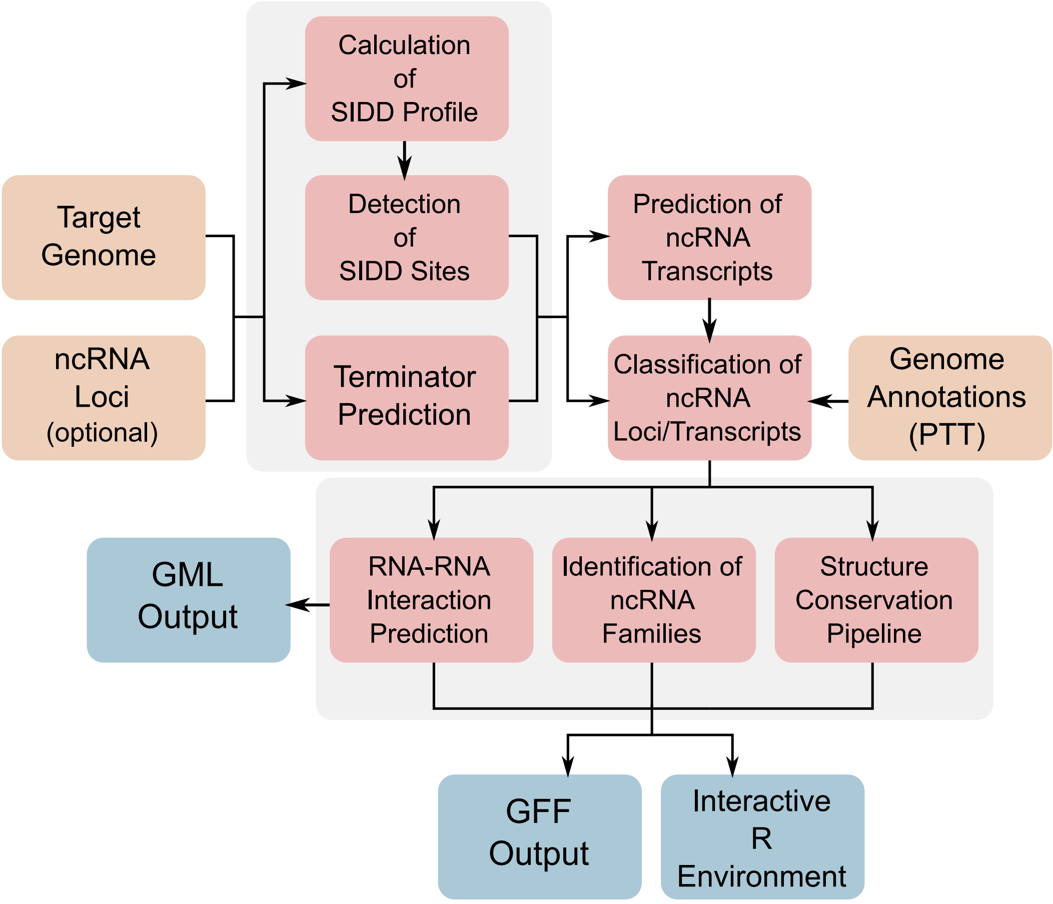

Calculation of SIDD Profile
Target Genome
Detection of SIDD Sites
Prediction of ncRNA Transcripts
ncRNA Loci (optional)
Terminator Prediction
Classification of ncRNA Loci/Transcripts
Genome Annotations (PTT)
GML Output
RNA-RNA Interaction Prediction
Identification of ncRNA Families
Structure Conservation Pipeline
GFF Output
Interactive R Environment

Figure 3.1.: Overview of NOCORNAc's workflow for the prediction and characterization of ncRNA transcripts. The target genome, genome annotations and predefined ncRNA loci (optional) are taken as input. Transcriptional features: A SIDD profile is calculated for the target genome and processed in order to detect SIDD sites, which are used to determine the 5' start of ncRNA transcripts. Rho-independent terminator signals are detected in order to define the 3' end of the transcripts. The detected signals are used to predict ncRNA transcripts and to classify ncRNA loci as transcripts or *cis*-regulatory elements. Characterization: The predicted ncRNA transcripts and ncRNA loci are further characterized, with respect to conserved secondary structures, potential RNA-RNA interactions and their assignment to known RNA families. Output: The results can be further assessed in NOCORNAc's interactive R environment. In addition, a GFF file comprising all results is generated. RNA-RNA interaction networks can be saved as GML files.

identified by RNAz. Any other method that is applicable for the detection of secondary structure conservation can be applied as well.

A second strategy is to omit the structure conservation analysis completely. In this case nocoRNAc combines SIDD sites and Rho-independent terminators to predict ncRNA transcripts in a genome-wide manner without considering regions of structural conservation (section 3.2.1). The combination of SIDD sites and terminator signals is restricted in this case. Features can only be combined if the resulting ncRNA transcript is not shorter than 30 bp and not longer than 600 bp. These thresholds can be customized.

The third and most sophisticated strategy starts with the genome-wide prediction of ncRNA transcripts and searches for conserved secondary structures for each candidate individually. This procedure is performed in several steps, which are combined in nocoRNAc's structure conservation pipeline (section 3.3).

3.1. Integrated Methods

3.1.1. Prediction of Rho-independent terminators

In most bacteria the transcription process of a large fraction of all transcripts is terminated by a so-called Rho-independent terminator [172]. These signals consist of a stem-loop structure followed by an A/T-rich region. The stem is usually built by a palindromic G/C-rich sequence, and thus it is quite stable. The size of the stem can vary but it is rarely longer than 20 bp. When the RNA polymerase reaches the terminator signal it interacts with the stem-loop structure which causes it to pause transcription. The A/T-rich region that follows the stem-loop allows the synthesized transcript to dissociate from the template, which also frees the polymerase and thus terminates transcription.

A second termination mechanism is facilitated by the Rho protein and is therefore called Rho-dependent transcription termination [127]. In this case the Rho protein binds to the so-called *Rho utilization site* on the synthesized RNA transcript. It then translocates along the transcript and reaches the polymerase, which is paused by a terminator signal at the end of the transcript that is usually quite similar to a Rho-independent terminator. There it acts as a helicase, which causes the RNA transcript to dissociate and the transcription process to be terminated. However, the Rho utilization site is often highly degenerated and hard to detect. Therefore, nocoRNAc focuses on the detection of Rho-independent terminator signals.

For this the program `TransTermHP` [82] is integrated in nocoRNAc. `TransTermHP` localizes the characteristic stem-loop motifs in bacterial genomes and assigns a score to each candidate which relates to the probability of the element to act as a Rho-independent transcription terminator. The scoring considers three different parts of the candidate, which are the stem, the loop and the tail. The tail is the single-stranded region downstream of the stem-loop. The scoring of the stem considers its size and the GC-content. The score of the loop is solely based on the size and the scoring of the tail considers the abundance of A/T nucleotides in its sequence as an A/T-rich tail leads to a less stable duplex between the transcript and the template leading to the dissociation of the transcript and thereby the termination of transcription. The combination of the three scores result in a single confidence value for each candidate.

3.1.2. Prediction of promoter regions: SIDD Sites

The computational detection of promoter regions or transcription start sites solely based on the genomic sequence is extremely challenging as there is a large number of sigma factors and other transcription factors that bind to certain signals in promoter regions and interact with the RNA polymerase to facilitate the initiation of the transcription process. The sequence motif of the transcription factor binding site varies significantly for the different transcription factors and in many cases the binding sites are short and can be highly degenerated or the sequence of the binding site is not known at all. Even when focusing on a specific organism the number of transcription factors that have to be considered can be very high. In *Streptomyces coelicolor*, for example, there are more than 60 known sigma factors [17].

Therefore, the prediction of ncRNA transcripts in a wide spectrum of bacterial genomes requires more general properties of promoter regions to be taken into account. A mechanism that is necessary for all sites of transcription initiation is the separation and partial unwinding of the DNA double helix at the respective locus to allow for the binding of transcription factors and the RNA polymerase. Thus, it can be assumed that the base composition in the context of a transcription initiation site potentially promotes this event.

A model that takes these features of promoter regions into account is the so-called SIDD model (Stress Induced Duplex Destabilization) [15]. This approach considers the thermodynamic stability of the base pairings on the dinucleotide level, the torsional energy that is needed for unwinding the helix and, in addition, the influence of superhelical stress.

The model was implemented as described in [15] and integrated in nocoRNAc to calculate a SIDD profile for a DNA sequence. The profile contains for each position the expected amount of free energy that has to be added to the system to establish a state in which the base pair at the respective position is separated. Each SIDD value of the profile is calculated on the basis of a partition function normalizing the free energy of all states in which the respective base pair is separated by the free energy of the whole system. For a region of length n there are 2^n possible different separation states and in theory the model would have to consider all of them to calculate the profile. This would result in an exponential time complexity and thus, only biologically plausible states are taken into account, which reduces the time complexity to $O(n^3)$. Biologically plausible states are characterized as states that contain no more than three continuous regions of separated base pairs.

For the calculation of a SIDD profile for a whole genome, a sliding window approach as suggested in [15] with a default window size of 5000 nt and a step size of 500 nt is used. With these settings each position is covered by ten windows and ten values are calculated, for which a weighted average is calculated. Here, a higher weight is assigned to values calculated for windows in which the position was located in the middle of the window whereas values near the border of a window receive a lower weight.

For maximal efficiency of the SIDD profile calculation only native Java arrays (`int`, `double`) have been used for its implementation. The calculation for a complete bacterial genome takes depending on the genome size a couple of hours and needs less than 512 MB memory (Tested on a Intel® Core™2 Quad Q9300 (2.5GHz)).

Parallelization To speed up the calculation process of the SIDD profile the procedure has been parallelized in this dissertation. For this, parts of the algorithm have been reimplemented to allow for a separate storing of the energy values of single states and the respective partition function which is used to calculate the Boltzmann factor for each state. Thus, the normalization using the partition function can be postponed to a later step and the windows can be calculated independently of each other. Then, after all windows have been calculated in parallel, the normalization is performed at once for the whole profile. The user can set the number of different cores/processors that are to be used for the calculation and the program splits up the windows in the respective number of different sets. The sets are then processed by independent threads, which, however, store there results in two global arrays. In one array for each genomic position the free energy values of all separations states are summed up that indicate a separation at that position. In the other array for each genomic position the energy values of all states are summed up that are covering the respective position. When the threads have finished the calculation the complete profile is finalized by normalizing the first array with the second one.

3.1.3. Conserved secondary structures: RNAz

Two of the three strategies of nocoRNAc to predict novel ncRNA transcripts are based on the detection of conserved secondary structures. One approach uses predicted regions of conserved secondary structure as a basis for the transcript prediction, the other approach, the structure conservation pipeline, applies this step after the prediction of a set of candidate loci.

For the task of predicting conserved secondary structures the program RNAz [61] is integrated in nocoRNAc. It was chosen because of its speed, robustness and good integratability. The RNAz program consists of the main method and supplementary scripts that are used for data preprocessing and postprocessing. However, only the core method is integrated as nocoRNAc is performing all data processing steps that are necessary. RNAz takes as input a set of aligned sequences.

Two properties are considered to predict if a given sequence alignment contains a functional structured RNA. The first property is based on the **minimum free energy** (MFE) of the RNA structure that is predicted for the given sequences. For this the program RNAfold is used [71]. The assumption is that a functional RNA with a structure on which its functionality depends has a significantly lower MFE, i.e., a more stable structure, than a random sequence. As a measure for how much the structure is more stable than one would expect by chance the z-score is used, i.e., the mean MFE as expected for a sequence of the same length and nucleotide composition is subtracted from the actual MFE of the query sequence and this is normalized by the respective standard deviation. The expected mean MFE and the standard deviation could be calculated by repeated shuffling and computational folding of the query sequence. However, RNAz uses a regression support vector machine (SVM) to estimate these values. The SVM was trained on different sets of sequences with varying lengths as well as mononucleotide and dinucleotide frequencies. Thus, for a given query sequence the regression SVM takes these properties as input and calculates the expected mean and standard deviation of the MFE, which is much faster as it could be accomplished by the shuffling approach. For each sequence in the alignment a z-score is calculated.

The second measure, which directly refers to structure conservation is the so-called **structure conservation index** (SCI). To determine the SCI the MFE structure is calculated for each sequence in the alignment using RNAfold [71]. In addition the MFE consensus structure for the complete alignment is calculated using RNAalifold [19]. The SCI is then determined by normalizing the MFE value of the consensus structure by the mean MFE value of the single structures. A SCI value close to 1 means that the MFE values of the single structures are very similar to the MFE of the consensus structure and this therefore indicates that the structure is well conserved. A low SCI indicates a low conservation of secondary structure. As a measure for secondary structure conservation the SCI has been shown to perform very well in comparison to other methods [60].

In the final classification step the mean z-score of all aligned sequences, the SCI as well as the mean pairwise identity (MPI) of the sequences are used as input for a classification SVM. This SVM was trained on Rfam alignments [29] as a positive set and on dinucleotide shuffled Rfam alignments as a negative set. The MPI has to be considered during the classification as very similar sequences are more likely to have a high SCI than more dissimilar sequences. The application of the classification SVM results in a P-value indicating the probability that the alignment contains a structured functional RNA. By default RNAz considers an alignment to be classified as an ncRNA if the P-value is equal to or greater than 0.5.

3.1.4. Identification of RNA families: Infernal

One relevant information about predicted ncRNAs is if they belong to an already known RNA family. Information about RNA families can be found in the Rfam database [29] and RNA family membership can be determined using tools such as `CMsearch` as described in section 2.2.3.

nocoRNAc integrates `CMsearch`, which is part of the `Infernal` package, to identify predicted ncRNAs that belong to annotated RNA families. For this, the Rfam database has to be provided to nocoRNAc in the form of an Rfam seed file, which can be downloaded from the Rfam FTP server. This file contains for each RNA family an alignment and a consensus secondary structure of family members in addition to supplementary information such as a description of the RNA family or suggested score cutoffs for the identification of new family members with `CMsearch`.

NOCORNAC uses the entries in the Rfam seed file to build covariance models for each RNA family by utilizing the tool `CMbuild`. Then `CMsearch` is applied to the target genome of NOCORNAC using these covariance models. The score cutoffs for the prediction are individually chosen for each RNA family as suggested in the respective Rfam seed entry. The results are collected to annotate predicted ncRNAs as members of known RNA families if possible.

The user can decide to search for all RNA families that are contained in the Rfam seed file or only a subset can be used by passing a list of the respective Rfam IDs to NOCORNAC. However, it is also possible to select certain categories of entries by passing the respective keywords (e.g., 'cis-reg' to only include *cis*-regulatory elements).

If a search is performed using the complete database, the procedure is very time-consuming and can take several days. For this reason the search process is parallelized in NOCORNAC and multiple queries can be processed at the same time.

3.1.5. RNA-RNA interactions: IntaRNA

Most known regulatory RNAs perform their function by base pairing with the mRNA of their target gene. Thus, the identification of potential target mRNAs of putative ncRNAs is a crucial step to predict regulatory function. Therefore, the program `IntaRNA` [30] is integrated in NOCORNAC to predict RNA-RNA interactions between putative ncRNAs and mRNAs of the bacterial organism to which NOCORNAC is applied. The advantage of `IntaRNA` is that it combines the energy of the duplex formation between the interacting RNAs but also the energy that is needed to unfold the RNAs such that the interaction site becomes accessible in both molecules. In addition, `IntaRNA` considers the existence of a seed region for the interaction of two RNA molecules. A seed region is a short stretch of nucleotides in both molecules that are complementary to each other and that are likely unpaired within the single molecules. Thus, the start of the duplex formation is favorable at these regions.

Most other programs for RNA-RNA interaction prediction are based on different approaches as summarized in [30]. One strategy is to concatenate the two candidate RNA sequences and predict the secondary structure of this merged sequence with a standard secondary structure prediction algorithm. Here, only small modifications of the algorithm are necessary to handle the part of the sequence where the two RNAs were joined. The disadvantage of these approaches is that they usually can only predict interactions, if the resulting joint structure is free of pseudoknots. However, many RNA-RNA interactions involve the loops of hairpin structures in the two RNAs, which would result in a joint structure with a pseudoknot. Another strategy is solely based on the free energy of the duplex formation. It could be shown that the additional consideration of the accessibility of the interaction site leads to a significant improvement of the prediction performance [30].

In `IntaRNA` the calculation of the hybridization energy of the duplex is based on the energy model used in `RNAhybrid` [126]. In order to consider the accessibility of the interaction site in both RNAs a partition function is used to calculate the difference between the energy of the ensemble of all secondary structures of the molecule and the energy of the ensemble of only those structures where the nucleotides of the interaction site are unpaired. The final energy of the predicted interaction is then defined as the sum of the hybridization energy of the duplex and the energy that is needed to make the interaction site accessible as calculated by the partition functions. An interaction between two RNAs is considered to be possible if this combined interaction energy is negative.

nocoRNAc provides two different approaches for RNA-RNA interaction prediction, for which `IntaRNA` is utilized. The first approach is aimed at the prediction of genome-wide networks of RNA-RNA interaction candidates. This strategy is described in section 3.4. The second approach is integrated in nocoRNAc's interactive R environment. It aims more at the target prediction for single pre-chosen elements and offers possibilities for the estimation of the significance of predicted interactions such as z-score and p-value calculation. This strategy is described in section 3.4.3.

3.2. Prediction of ncRNA transcripts

The first strategy that nocoRNAc can employ for the detection of ncRNA transcripts is based on ncRNA candidate loci, which, for example, result from a genome-wide application of a program such as RNAz [61], that identifies regions with conserved secondary structures. As a first step nocoRNAc annotates these loci with all transcriptional features that have been predicted in the respective region. For the transcript prediction SIDD sites and terminator signals are combined, which allows for a determination of the strand of the transcript and a localization of transcript boundaries that is more precise as by the RNAz prediction alone.

The prediction algorithm takes each SIDD site into consideration that has been predicted within an ncRNA candidate locus or not further than 25 bp apart. For each site a matching terminator is searched in both directions as SIDD sites are not strand-specific. The start of the candidate transcript is set to the coordinates of the SIDD site and as the end the first high-confidence terminator downstream or upstream of the SIDD site is selected. A predicted high-confidence terminator has a confidence of 75 or greater [82]. If no high-confidence terminator can be found, the predicted transcript is extended to the terminator with the highest confidence value. If there were no terminator signals predicted for the respective candidate region, the boundaries of the region itself are used to terminate the predicted transcript. However, transcripts without a predicted terminator are only kept, if their SIDD site is not located in the upstream region of a protein-coding gene, as in this case the site is potentially associated to the transcription start of that gene.

After the prediction of candidate ncRNA transcripts overlapping predictions are merged, if they are located on the same strand. For overlapping transcripts on different strands only the candidate with the higher terminator confidence is kept while the other candidate is shortened by selecting an alternative terminator signal that is closer to the respective SIDD site. If no alternative terminator can be found, the candidate transcript is completely removed from the predictions.

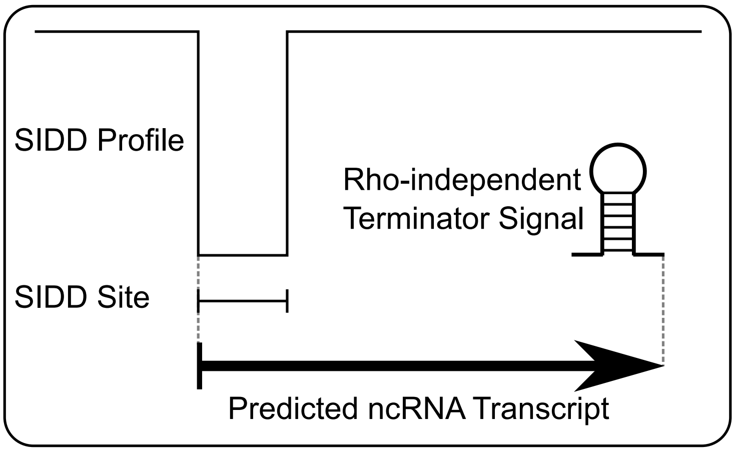


Figure 3.2.: nocoRNAc's ncRNA prediction concept. SIDD sites, which are drops in the genomic SIDD profile, indicate regions where an opening of the DNA duplex is favorable. Predicted ncRNA transcripts start at a SIDD site and are extended to the best Rho-independent terminator signal or the first high-confidence terminator signal which is found downstream of the SIDD site.

3.2.1. Genome-wide transcript prediction

nocoRNAc also allows for a prediction of ncRNA transcripts independently of predefined ncRNA regions. This is useful if there is no multiple genome alignment available to apply RNAz or similar methods for the detection of ncRNA regions or if the user wants to detect transcripts without any predefined constraints with respect to certain loci. The genome-wide transcript prediction uses the same approach as in the context of ncRNA regions (see above), but here is is applied to the complete chromosome. This has certain implications. First, the prediction of transcripts without terminator is not possible anymore. In the context of a predefined region the end of that region can be set as the transcript end if no terminator signal can be found. In the context of the whole genome this is not possible. In addition, the prediction procedure potentially considers all terminator signals downstream or upstream of a SIDD site as a possible 3' end of the transcript. This is also not feasible in a whole-genome context. Therefore, only terminators with a maximal distance of 600 bp downstream of the TSS are considered by default.

A schematic representation of the prediction concept is depicted in figure 3.2.

If predefined ncRNA regions are used as input, this usually means that there is already a certain indication that these loci contain ncRNAs, e.g., if they have been determined by RNAz. However, in the genome-wide approach only transcriptional signals are considered for the transcript prediction and thus, the predictions might also contain small proteins, which have not been annotated. In order to collect additional evidence that the predicted loci contained structured RNAs, nocoRNAc's structure conservation pipeline is applied (see section 3.3).

3.2.2. Integration of nocoRNAc with TSS prediction

Instead of using predicted SIDD sites to determine the start positions of putative ncRNA transcripts nocoRNAc is able to utilize the results of a comparative TSS prediction on RNA-seq data (see chapter 6). For this the TSS prediction algorithm can optionally provide the results in a format that can be directly read by nocoRNAc. However, the 3' end of the transcript is still determined by the prediction of Rho-independent terminator signals. In contrast to the standard prediction of ncRNA transcripts each transcript start is considered to be real even if a matching terminator signal cannot be found. In such a case the end of the transcript is assumed to be located 150 bp downstream of the TSS. This is a default value that was chosen according to the properties of typical sRNAs. Depending on the target organism or the type of ncRNA that shall be identified this value can be adapted. It should be noted that the TSS, which are used as input for nocoRNAc do not have to be filtered according to their classification (*Primary*, *Secondary*, *Internal* or *Antisense*; see section 6.1.4) as nocoRNAc automatically classifies the transcripts according to their location relative to protein-coding genes and also prevents sense-overlapping transcripts from being predicted. Thus, the classification parameters, e.g., the assumed UTR length, are not relevant for nocoRNAc.

3.3. nocoRNAc's structure conservation pipeline

A conserved RNA secondary structure is one of the most important criteria of ncRNA prediction. For the assessment of secondary structure conservation in predicted ncRNA transcripts a structure conservation pipeline was integrated in nocoRNAc. The pipeline is able to automatically collect for each candidate a set of sequences with an optimal evolutionary distance from a sequence database. These sequences are then aligned and further preprocessed before they are used as input for RNAz. This allows for an assignment of RNAz P-values to ncRNA candidates without the necessity to decide on related organisms for a whole genome alignment. As all steps of the pipeline depend on the previous ones they are applied sequentially for each candidate. However, the processing of the set of candidates is parallelized. A schematic representation of nocoRNAc's structure conservation pipeline is shown in figure 3.3.

3.3.1. Collecting homologous sequences

RNAz is applied to a set of aligned sequences that have to be in an appropriate evolutionary distance to each other in order to predict if they share a conserved secondary structure. An optimal mean pairwise identity (MPI) of the alignment is around 80% [61]. Alignments with an MPI significantly above 90% are problematic as sequences that are nearly identical naturally fold into almost the same structure. Therefore, reliable information about conserved structures can only be derived from divergent sequences. An MPI below 50% should also be avoided as an accurate prediction of the consensus structure becomes infeasible.

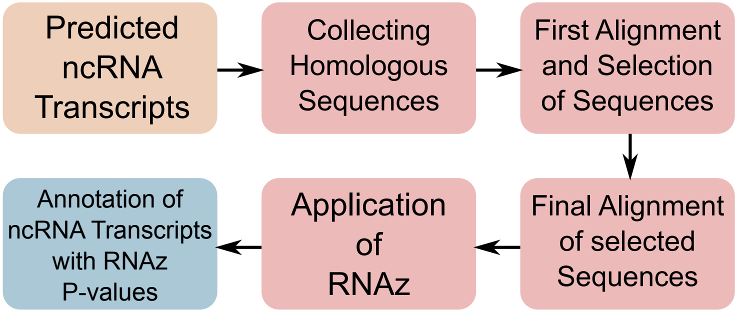

Predicted ncRNA Transcripts
Collecting Homologous Sequences
First Alignment and Selection of Sequences
Final Alignment of selected Sequences
Application of RNAz
Annotation of ncRNA Transcripts with RNAz P-values

Figure 3.3.: Overview of NOCORNAC's structure conservation pipeline. For each predicted ncRNA transcript homologous sequences are collected using BLAST. The sequences are aligned with ClustalW and sequences with an optimal pairwise identity are selected. A final alignment is generated for the selected sequences and RNAz is applied to the final alignment. The predicted ncRNA transcripts are finally annotated with the RNAz P-values, which indicate the probability of a conserved secondary structure.

Considering these constraints the structure conservation pipeline has been designed to collect a set of sequences for each ncRNA candidate individually whose MPI is as close as possible to the optimum of 80%. Thus, as a first step a nucleotide BLAST search (`blastn`) [6] in a sequence database is performed for the ncRNA candidate. This database can be for example the NCBI "Nucleotide collection (nt)", which can be downloaded from the NCBI FTP server[1]. However, as this database includes also eukaryotic sequences it is more efficient to construct a custom database from all bacterial genome sequences, which can also be downloaded from NCBI[2].

All BLAST parameters are customizable in NOCORNAC's configuration file. By default standard `blastn` parameters are used. An exception is a reduction of the word size to 7, which is necessary to also be able to align short sequences. In addition to that the filter for low complexity regions is switched off, as it appeared that this filter decreases the overall sensitivity especially for ncRNA candidates containing repetitive regions such as the poly-A region following many Rho-independent terminator signals.

3.3.2. Preliminary alignment and selection of sequences

All hits with a pairwise identity in comparison to the query below 50% and above 95% are discarded. If the number of remaining sequences is insufficient the ncRNA candidate is not further processed and no RNAz P-value is computed. By default the minimal alignment size is 2. The accepted hits are aligned using ClustalW [154] to determine the best set of sequences for the final alignment that is used as input for RNAz. Starting with the sequence that is closest to the optimal pairwise identity of 80% to the query, sequences are iteratively added to the final alignment set always

[1] ftp://ftp.ncbi.nih.gov/blast/db/

[2] ftp://ftp.ncbi.nih.gov/genomes/Bacteria/all.fna.tar.gz

selecting the sequence which leads to the MPI closest to 80%. This is done until the maximum alignment size is reached, which is 4 by default.

3.3.3. Final alignment and P-value calculation

The selected sequences are then realigned using ClustalW and the resulting final alignment is used as input for RNAz. The ncRNA candidate is annotated with the resulting P-value. If the length of the alignment exceeds 200 bp it is sliced by a sliding window approach with a window size of 200 bp and an offset of 20 bp. Each alignment window is separately analyzed with RNAz. The ncRNA candidate is then annotated with the best P-value of all windows. Therefore, a predicted transcript gets a positive prediction even if only a part of it is potentially structurally conserved. This increases sensitivity for longer transcripts that also contain regions without any conserved structure. In addition, the sliding window approach reduces the run-time significantly.

3.3.4. Application to kingdom-wide ncRNA prediction

In a study conducted by Andreas Friedrich in his master thesis [49] the structure conservation pipeline of nocoRNAc was applied to the prediction of ncRNA transcripts in the genomes of 125 bacteria from 25 different phyla. This resulted in altogether 68, 643 predictions, where the number of putative ncRNAs per genome varied between 82 and 1633. The vast majority of transcripts was predicted antisense to protein-coding genes. All predicted elements were further characterized with respect to their conservation in other species and several other properties. In addition, the `GraphClust` pipeline [70] has been applied to cluster ncRNA candidates with respect to their secondary structure.

It turned out that most of them are only conserved on the species level with only a few being found across phyla, most of which representing housekeeping RNAs like rRNAs or tRNAs. Interestingly, some elements have been identified that are also found in only a few species but they belong to different phyla. A more detailed analysis showed that in these cases the respective species usually populate similar habitats. Horizontal gene transfer between divers bacteria that can be found in similar habitats or host environments has been shown to be quite likely [123]. Thus, the elements identified here might also have been subject to horizontal gene transfer. More thorough studies are required in this context in order to rule out that the effect is due to a bias in the prediction procedure or due to the chosen thresholds.

Another observation that has been made is that *cis*-asRNA candidates tend to have significantly lower RNAz P-values in comparison to predicted intergenic ncRNAs. A reason for this might be the slightly different mechanisms of regulation. Both, *trans*-encoded and *cis*-encoded regulatory RNAs regulate their target by RNA-RNA interaction with its mRNA. However, for *cis*-encoded ncRNAs, which are located antisense to their target gene, the size of the complementary region is often very large, while the interaction sites in *trans*-encoded ncRNAs is shorter. Thus, for *trans*-encoded elements the secondary structure potentially plays a more important role, which results in more conserved structures for this group of ncRNAs.

This observation shows that the prediction of *cis*-encoded ncRNAs is more challenging than for intergenic elements. In addition to the low structure conservation, sequence conservation is a problematic criterion as a search for homologous sequences is biased by the sequence of the protein-coding gene that is located antisense to the RNA. Thus, a combination of computational methods and experimental results as presented in chapter 7 is necessary to enhance the performance of asRNA detection significantly.

3.4. Prediction of RNA-RNA interaction networks

nocoRNAc is able to automatically apply the RNA-RNA interaction prediction program `IntaRNA` to a set of predicted ncRNA regions and protein-coding genes that are selected by the user. The results are filtered and included in the automatic annotation of predicted ncRNA regions. The filtering is done with respect to the free energy and to the size of the interacting regions. The respective thresholds can be customized by the user. By default they are dynamically chosen by nocoRNAc so that only the n best percentiles of all interactions are considered. n is set to 2 by default but this value can also be adjusted in the configuration file. Interactions contained in the best percentile are regarded as high-scoring interactions with respect to their energy value and/or length.

The resulting interaction network is also used to generate a GML or DOT file that can be visualized by a graph visualization tool. In this network, nodes represent the interacting elements, where ncRNAs are shown as squares and protein-coding genes are shown as circles. Edges represent predicted interactions. Interactions selected because of their low energy value are shown in red whereas interactions selected because of their length are shown in blue. An example of a small interaction network is shown in figure 3.4.

3.4.1. Transcript Interaction Profiles

A functionality of nocoRNAc related to the prediction of RNA-RNA interactions is the calculation of *interaction profiles* for each RNA that is involved in at least one predicted interaction. The interaction profile is a graph that assigns to each nucleotide position of the respective RNA a value that describes its participation in predicted interactions.

More precisely there are 4 different types of interaction profiles. $n(x)$ denotes the number of interactions nucleotide x is involved in. $g(x)$ denotes the expected free energy value of an interaction in which nucleotide x is involved. $p(x)$ denotes the probability that nucleotide x participates in an interaction in which its RNA is involved and $p_{net}(x)$ denotes the probability that nucleotide x is involved in any interaction considering all interactions in the network. These profiles are calculated as follows.

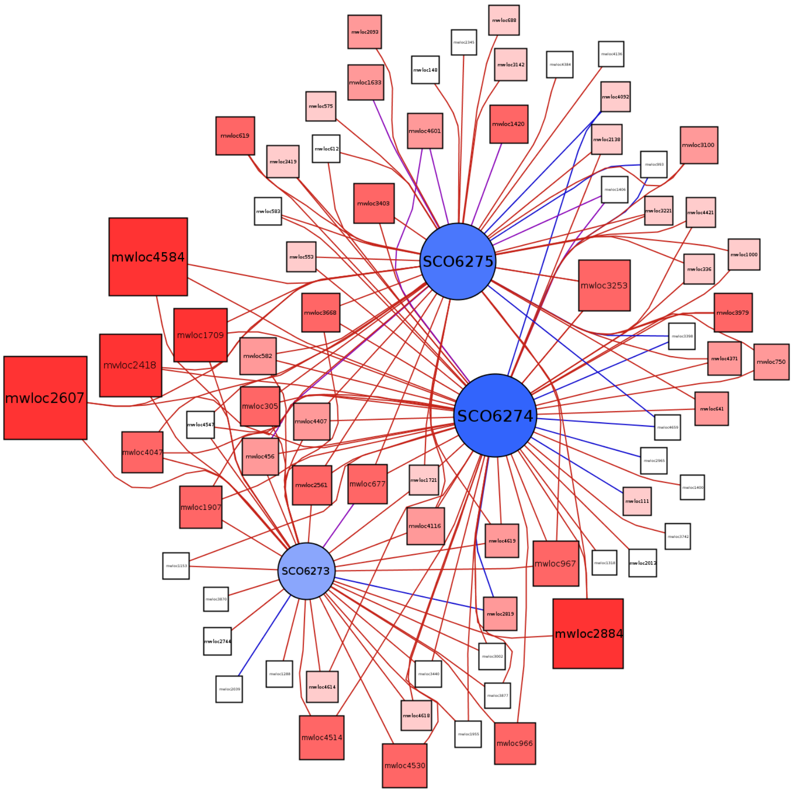


Figure 3.4.: Example of an RNA-RNA interaction network. Protein-coding genes are depicted as blue circles, ncRNAs are depicted as red rectangles. The degree of the nodes is denoted by their size and their opacity. Red edges show high-scoring interactions with respect to the free energy value, while blue edges denote interactions with long interaction sites. Interactions shown as purple edges have a good free energy value and a long interaction site. Interactions between a protein-coding gene and its *cis*-encoded asRNA are shown in black. The network is visualized with `yEd` [180].

Let s be a single RNA sequence and $s(x)$ the xth nucleotide of that RNA. I_s is the set of interactions in which s participates and $I_{s(x)}$ is the subset of such interactions that involve the xth nucleotide of s. Then the four profiles are defined as:

$$n(x) = |I_{s(x)}|, \quad g(x) = \frac{\sum_{i \in I_{s(x)}} E_i e^{-\frac{E_i}{RT}}}{\sum_{i \in I_{s(x)}} e^{-\frac{E_i}{RT}}}, \tag{3.1}$$

$$p(x) = \frac{\sum_{i \in I_{s(x)}} e^{-\frac{E_i}{RT}}}{\sum_{i \in I_s} e^{-\frac{E_i}{RT}}}, \quad p_{net}(x) = \frac{\sum_{i \in I_{s(x)}} e^{-\frac{E_i}{RT}}}{\sum_{i \in I} e^{-\frac{E_i}{RT}}}, \tag{3.2}$$

where E_i is the free energy value of interaction i, R is the gas constant, T is the temperature in Kelvin and I is the set of all interactions in the network. $g(x)$ is set to zero, if nucleotide x is not involved in an interaction.

Using these formulas, the contribution of an interaction to a profile is weighted with its Boltzmann factor except for the profile $n(x)$. The profile types have different applications. They can all be used to identify regions of an RNA that are likely to be involved in an interaction. p can be used to compare regions within the same RNA, whereas p_{net} is more suitable for inter profile comparisons.

An example of an interaction profile plot is shown in figure 3.5.

3.4.2. Transcript Interaction Matrix

To allow for a quick evaluation of the most probable partners of each RNA contained in the predicted interaction network the *transcript interaction matrix* (I_{mat}) is calculated. It denotes for each element x and any other element y the probability that the interaction partner of x is y. A single cell (x, y) of the matrix is calculated as follows:

$$I_{mat}(x, y) = \frac{\sum_{i \in I_{x,y}} e^{-\frac{E_i}{RT}}}{\sum_{i \in I_x} e^{-\frac{E_i}{RT}}}, \tag{3.3}$$

where E_i is the free energy value of interaction i, R is the gas constant, T is the temperature in Kelvin, I_x is the set of all interactions element x is involved in and $I_{x,y}$ is the set of all interactions between x and y. Note that $|I_{x,y}|$ is usually 0 or 1 as only the optimal interaction for a pair is predicted, if default settings are used.

The resulting matrix can be visualized as a heatmap, for example.

3.4.3. Interactive RNA-RNA interaction prediction in nocoRNAc's R environment

A more customizable functionality for RNA-RNA interaction prediction is integrated in nocoRNAc's R environment. Here, the function `intarna` can be used to predict interactions between individual RNAs or sets of RNAs that are available in the

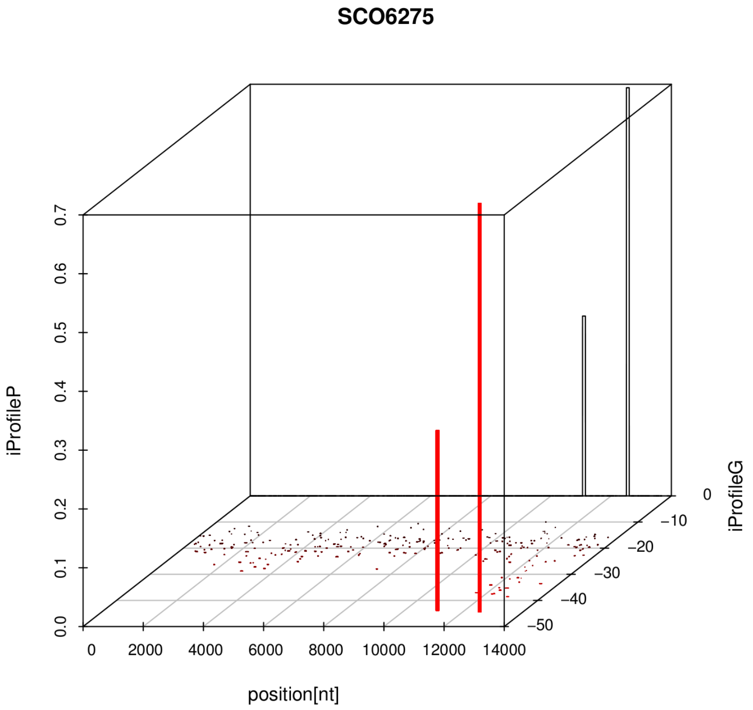


Figure 3.5.: Example of an RNA-RNA interaction profile plot showing position, probability and free energy of all interactions predicted for a protein-coding RNA. In this example the 3' end of the RNA has two high-scoring interactions.

environment. This includes ncRNA regions (as for example calculated by RNAz), ncRNA transcripts predicted by NOCORNAC and gene annotations. In addition to that users can import their own annotations or create annotations in the environment. It is also possible to enter sequence information directly.

This functionality is therefore extremely flexible allowing for the prediction of interactions between any kind of RNA molecules, which is not limited to interactions between ncRNAs and mRNAs. Interactions between two ncRNAs or two mRNAs can also be predicted, for example.

Most sRNAs regulate the translation of an mRNA by binding to its 5' UTR and sequestering the ribosome binding site. Although examples are known where the sRNA binds to the coding region of the mRNA, it is in general desirable to include the 5' and 3' UTRs of annotated genes in the RNA-RNA interaction analysis. For that reason it is possible to specify an upstream and downstream context size that is additionally considered during the interaction prediction.

The most important property of predicted RNA-RNA interactions that can be used for their evaluation is the predicted free energy of the interaction. This value, which is provided by `IntaRNA` includes the energy that is needed to make the interaction sites in both molecules accessible and the free energy of the duplex formation between the two molecules. The interaction can only be assumed to take place if this value is negative.

However, it is generally difficult to estimate how low the free energy value has to be to assume a strong and highly probable interaction. This is because this value is influenced by various properties of the RNA molecules such as the length of the two sequences and their GC-content. Therefore, it is useful to evaluate the significance of the interaction statistically instead of relying on the free energy value alone.

For this the `intarna` function in NOCORNAC's R environment incorporates methods for z-score and p-value calculation. In order to calculate these values the sequences of the RNAs that are subject to the interaction prediction are shuffled several times and each time the interaction prediction is repeated. By default only the target sequence is shuffled as the standard application of the `intarna` function is to search for possible targets of an ncRNA among all protein-coding genes. z-scores are then computed by subtracting the mean free energy value of these samplings from the free energy value of the original prediction and dividing by the standard deviation. Two approaches are used to calculate p-values. In the first approach the original free energy value is subtracted from all values of the samplings. Then a one-sided one-sample t-test is performed on the resulting values (alternative hypothesis: true mean is greater than 0). In a second approach the fraction of sampled values that is greater or equal to the original predicted value is computed. By default the first approach is used as for the second approach the number of samplings that is needed to estimate reliable p-values is much higher, which especially affects small p-values.

Using z-scores and p-values in addition to the free energy values allows for much more precise evaluation of the predictions that is independent from sequence length and composition.

3.5. nocoRNAc's interactive R environment

nocoRNAc produces a single output file in GFF format, which contains all predicted features (SIDD sites, terminators), ncRNA transcripts and also protein-coding regions. More detailed information on these predictions is provided as attributes to the respective entries. This file can be used as input for genome browsers, for example.

However, the GFF file represents a very static representation of the prediction results. Depending on the analysis the user wants to perform on the data, several postprocessing steps such as filtering might be necessary. These analyses steps have to be performed dynamically in many cases, i.e., thresholds for subsequent filtering steps might depend on the results of the previous ones, which make continuous user interaction necessary. Additionally, the user might want to investigate interim results by means of statistics or visualization.

To allow for this dynamic analysis of the results nocoRNAc is able to provide parts of its data structure within an interactive `R` [125, 53] environment, allowing the user to perform a variety of statistical analyses to the results as well as to visualize them.

Most of the data structures are presented as a `data.frame`. All tables listing information about sequence features, i.e., regions in a genomic context, contain at least the columns `start`, `end` and `strand` to define the respective locus. This applies to `sidd.sites`, `terminators`, `nc.transcripts`, `ncRNAs` and `genes`. Single or multiple lines of these tables can be used as input for the function `getSequences` which returns a set of respective DNA sequences, to which all functions of the Bioconductor package `Biostrings` [113] are applicable, thus allowing for sequence manipulation or conversion (e.g., translation to AA sequences) or export as multi FASTA file.

The elements can also serve as input to the `intarna` function for the prediction of RNA-RNA interactions (see 3.4.3).

ncRNA transcripts and related features The `sidd.sites` and `terminators` tables contain energy values and confidence scores of all predicted SIDD sites and terminator signals, respectively. The `nc.transcripts` table contains information about predicted ncRNA transcripts such as the identifier of the ncRNA region in whose context the transcript was predicted as well as the identifiers of the SIDD site and the terminator, by which the transcript is defined. Finally, each entry contains a list of all protein-coding genes to which the ncRNA transcript is located antisense.

ncRNA regions The `ncRNAs` table, which represents the set of ncRNA regions, which have been predicted by RNAz, has a different kind of structure, i.e., it is a list of lists. Each element contains the entries `$start`, `$end`, `$strand` so that it can be handled by the `getSequences` and the `intarna` function. Additionally it contains the RNAz P-value (`$score`), nocoRNAc's classification information (`$class`, lists of all SIDD sites and terminators that have been assigned to that region (`$sidd.sites`, `$terminators`), a list of all protein-coding genes that are overlapping that region

(`$genes`) and a list of all ncRNA transcripts that have been predicted in the context of that region (`$pred.transcripts`).

genomic sequence and SIDD profile In addition to these data structures NOCORNAC's R environment provides the complete genome of the target organism (`genome`) as a `DNAString` object, which serves as the basis of the `getSequences` function but which can also be freely accessed by the user with all functionalities of `Biostrings`.

The SIDD profile, which is calculated by NOCORNAC for the target genome, is provided as a numeric vector `sidd.profile` with the length of the genome and a SIDD value for each genomic position. Therefore, the SIDD profile can be visualized for any genomic region by using standard plotting functions implemented in R.

RNA-RNA interactions If an RNA-RNA interaction network has been computed the respective results are also provided in NOCORNAC's R environment. The primary information about all predicted interactions is listed in the table `interactions`. For each interaction this `data.frame` contains the IDs of the transcripts that are involved in the interaction, the coordinates of the interaction region in relation to the transcripts and the free energy value of the interaction. In addition, the individual energy values are listed which are combined to calculate the free energy value of the interaction. I.e. the energy that is needed to make the interaction site accessible for each of the transcripts and the hybridization energy of the duplex.

The interaction profiles that have been calculated for each transcript (section 3.4.1) are provided as numeric matrices. These matrices are `iProfileN` ($n(x)$), `iProfileP` ($p(x)$), `iProfilePnet` ($p_{net}(x)$) and `iProfileG` ($g(x)$).

These profiles can be visualized individually for each transcript or as multi profile plots for a set of transcripts using R's standard plotting function. To visualize a combination of $p(x)$ and $g(x)$ as a 3D plot NOCORNAC's R environment integrates the function `iProfilePGplot`. It takes the ID of an interacting transcript as input and visualizes the two respective profiles by the help of R's `scatterplot3d` [91]. This allows for an easy identification of all relevant interaction sites of the transcript with an indication of their probability and free energy value. An example for such a visualization is shown in figure 3.5.

The interaction matrix containing all pairwise interaction probabilities of all transcripts (section 3.4.2) is provided as a numeric matrix (`iMatrix`). This matrix or parts of it can for example be visualized as a heatmap. It can also be used to determine the most probable interaction partners of a transcript or a group of transcripts.

4. ncRNAs as regulators in the model bacterium Streptomyces coelicolor

Streptomyces coelicolor is a Gram-positive antibiotics producing soil bacterium and a model organism of the genus *Streptomyces* [17, 115]. Its linear chromosome is more than 8Mb in length, has a high G/C content (72%) and encodes about 7800 genes. The life cycle of *S. coelicolor* undergoes a metabolic switch from primary metabolism in the exponential growth phase to secondary metabolism during stationary growth. The secondary metabolism is characterized by the production of, so-called, secondary metabolites. Many of these substances have an antibiotic effect. Most prominently the antibiotics actinorhodin (ACT), undecylprodigiosin (RED), and the calcium-dependant antibiotic (CDA) are produced by *S. coelicolor*. However, many more secondary metabolites have been identified, most of which are only produced under very specific conditions and often in very low concentrations. Furthermore, their function remains unknown in many cases. In addition, several, so-called 'cryptic' secondary metabolic pathways have been identified, which are predicted to be involved in the production of secondary metabolites that have not been measured yet. As the produced antibiotics could be clinically relevant the mechanisms involved in the regulation of secondary metabolism are of major interest and turned out to be very complex [20].

In the SysMO-STREAM consortium the transcriptome of *S. coelicolor* undergoing the metabolic switch has been studied in unprecedented detail [109, 164, 99]. For this, *S. coelicolor* wild type and mutant strains were grown under various nutrient limiting conditions during controlled submerged batch fermentations and a custom design microarray including probes for intergenic regions and predicted ncRNAs was used to produce transcriptomic time series data in a highly reproducible manner [13]. These analyses were complemented by proteomic and metabolomic studies [153, 3]. The aim of these studies was to elucidate regulatory mechanisms controlling the metabolic switch and the production of secondary metabolites.

In the context of regulation non-coding RNAs (ncRNAs) are of increasing interest [12] (see also section 2.1). In bacteria this most importantly involves mechanisms related to pathogenicity [121, 55, 52, 1], specific housekeeping functions or adaptation to various stress situations [106, 182, 144]. In order to study the role of ncRNAs in the regulation of the metabolic switch and secondary metabolism the ncRNA transcript prediction and characterization program nocoRNAc (chapter 3) was applied to the genome of *S. coelicolor* [65]. The results of this genome-wide ncRNA prediction are presented in section 4.1. Transcriptomic data from a time series expression analysis of *S. coelicolor* wild type grown under phosphate limited conditions [109] was used to validate and characterize the expression of the predicted elements (section 4.2). Furthermore, a comparative gene expression analysis in *S. coelicolor* wild type and a SC*glnK*-3 mutant strain under glutamate limited conditions is presented in section 4.5 [164]. Sections 4.3 and 4.4 describe the prediction of RNA-RNA interactions between putative ncRNAs and protein-coding genes, which led to the identification of putative ncRNA transcripts potentially regulating antibiotics production in *S. coelicolor*.

4.1. Genome-wide prediction and characterization of ncRNAs in S. coelicolor

The first step of the analysis was the genome-wide prediction of candidate ncRNA loci in *S. coelicolor*, for which the program `RNAz` [61] was used. `RNAz` needs a multiple sequence alignment as input, which is then classified as potentially containing a conserved structured RNA or not. The sequence alignments used as input were gathered as follows. As the basis a whole-genome alignment of the three related *Streptomyces* species *S. coelicolor* [17], *S. avermitilis* [75] and *S. griseus* [112] was generated using the genome alignment software `Mauve` [37, 38]. For further processing the `xmfa` alignment as produced by `Mauve` was converted into `maf` format.

`RNAz` applies its classification procedure to the complete alignment which is used as input. Thus, the whole-genome alignment needs to be sliced prior to `RNAz` application. In order to be able to detect ncRNA transcripts of different sizes a sliding window procedure using different window sizes was applied to the alignment. Here, window sizes of 60, 80, 100, 120 and 160 nucleotides and a step size of 20 nucleotides were used. `RNAz` was applied to all windows that contain alignment information for all three species. Using these criteria 34.6% of the genome of *S. coelicolor* was covered. An `RNAz` SVM P-value of 0.5 has been used as threshold to classify an alignment window as potentially containing a conserved structured RNA. All overlapping windows with a positive classification were merged and used as input for NOCORNAC for further processing.

The `RNAz` application resulted in the prediction of 4,707 ncRNA loci in the genome of *S. coelicolor*. After processing with NOCORNAC 2,358 of these regions were annotated with a Rho-independent terminator signal and 2,237 regions were annotated with a SIDD site. The application of NOCORNAC's ncRNA transcript prediction procedure to annotated regions resulted in 843 putative ncRNA transcripts. Comparison with genome annotations showed that 653 predicted transcripts are antisense to protein-coding regions while 180 are located intergenic and 10 are partially overlapping a protein-coding region on the sense strand. The predicted elements were also compared to annotated known ncRNAs like rRNAs, tRNAs, etc. This showed that 96 of the intergenic predictions correspond to known ncRNAs while 84 are putatively novel ncRNA transcripts.

NOCORNAC's interactive `R` environment allows for a more detailed analysis of the whole dataset or single elements. In addition, a functionality for the visualization of ncRNA loci and related information is provided. Example plots for two 5S ribosomal RNAs are shown in figure 4.1. In addition to the predicted transcript, the plots contain the ncRNA locus as predicted by RNAz and transcriptional features like

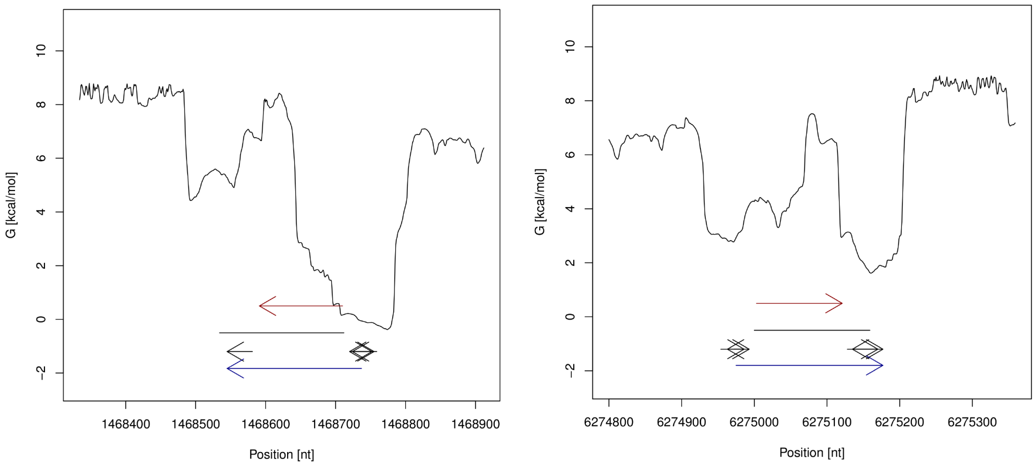


Figure 4.1.: Transcription feature plots of ncRNA transcripts predicted by nocoRNAc (blue arrows) covering annotated ribosomal RNAs (red arrows). The SIDD profile of the genomic region is drawn as a black graph (related scale on the y-axis). The coordinates of the genomic region are denoted on the x-axis. The ncRNA locus predicted by RNAz is shown as a black line. Predicted Rho-independent terminator signals are depicted as short black arrows. nocoRNAc considers the properties of the predicted transcription features (free energy value of SIDD sites; confidence value for terminators) and not only their position to predict the strand.

Rho-independent terminator signals and the SIDD profile for the entire region. With the help of these plots specific predictions can be evaluated, e.g., with respect to all signals in the region, which might give rise to alternative predictions.

Each predicted ncRNA transcript starts at a so-called SIDD site, which is defined as a genomic locus where a significant drop in the SIDD profile is observed. These SIDD sites can have a length of up to several tens of base pairs. This makes a precise localization of the transcription start difficult. To increase the probability of including the whole transcript in the prediction the start of the transcript is predicted at the start of the SIDD site. Note that the transcript might actually start further downstream. The localization of the terminator signals is more precise, but in many cases there are several signals predicted for a region. This can also be observed for the examples shown in figure 4.1. In both cases the actual transcripts end some base pairs further upstream than the prediction. It also has to be noted that nocoRNAc includes the whole terminator in the prediction while it is usually not considered to be part of the respective RNA motif as annotated in the database (e.g. Rfam).

In many situations there are transcriptional start and also termination signals at both sites of the ncRNA locus as can be seen in figure 4.1 (right). This makes the determination of the strand of the ncRNA transcript difficult. In such a case nocoRNAc selects the transcript with the stronger signals, which allowed for the correct prediction of the strand at the depicted loci.

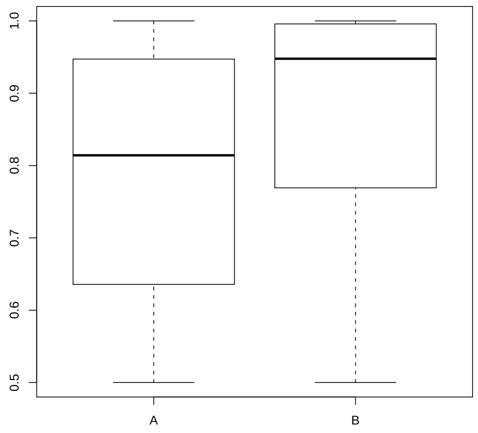


Figure 4.2.: Boxplots of RNAz P-value distributions of candidate loci in *S. coelicolor* without a transcript predicted by NOCORNAC (A) and regions for which an ncRNA transcript was predicted (B).

The degree of structural conservation as measured by RNAz is one of the most important criteria in the context of ncRNA prediction. As the primary function of NOCORNAC is to improve ncRNA predictions by integrating transcriptional feature detection, a relevant question is if there is a correlation between the RNAz P-value and the strength of transcriptional features of the candidate loci. To answer this question the predicted ncRNA loci of *S. coelicolor* were grouped in two sets. One set (A) consists of loci for which no ncRNA transcript was detected by NOCORNAC. The second set (B) contains candidate loci with a transcript prediction. The P-value distribution of the loci containing a transcript has a significantly higher mean than the respective distribution of the loci without a transcript prediction (figure 4.2). The P-value of more than 60% of the regions for which a transcript was predicted exceeds 0.9. Furthermore, a one-sided two-sample T-test that was applied to the two distributions rejected the null hypothesis with a p-value of $6.66 \cdot 10^{-49}$. The effect gets even more pronounced if transcripts are only predicted with high confidence signals. Stricter thresholds of 4 kcal/mol for the SIDD site detection and 76 for the terminator confidence resulted in more than 90% of the transcript containing candidate loci having an RNAz P-value exceeding 0.9.

To evaluate NOCORNAC with respect to sensitivity and specificity the ncRNA candidate loci as predicted with RNAz and the ncRNA transcripts predicted by NOCORNAC were compared to all annotated ncRNAs for *S. coelicolor* in NCBI Genbank [16] and the Rfam database (10.0) [50]. The results of this evaluation are summarized in table 4.1.

RNAz predicted an ncRNA candidate locus for all 21 annotated ncRNA genes (not considering tRNAs), of which NOCORNAC classified 16 (76%) correctly as ncRNA transcripts. For this evaluation a correct classification was only counted if the strand of the transcript was correctly predicted by NOCORNAC. If the strand information

Table 4.1.: Comparison of predicted ncRNA loci and transcripts to annotation from NCBI and Rfam for *S. coelicolor*.

annotated ncRNAs	**RNAz locus**	**predicted transcript nocoRNAc [correctness %]**	**predicted transcript SIPHT [correctness %]**
21 ncRNA genes	21 (100%)	16 (76%)	13 (62%)
65 tRNAs	57 (88%)	30 (53%)	1 (2%)
28 *cis*-regulatory motifs	17 (61%)	1 (94%)	2 (93%)

The first column contains the numbers of annotated elements for 3 types of ncRNAs in *S. coelicolor*: ncRNA genes (without tRNAs) and tRNAs from NCBI as well as *cis*-regulatory motifs from Rfam. The second column indicates the number of elements for which RNAz predicted an ncRNA locus (strand-unspecific). Columns 3 and 4 indicate the number of annotated elements predicted to be an ncRNA transcript (strand-specific) by nocoRNAc and SIPHT, respectively.

is disregarded, nocoRNAc finds 19 of 21 transcripts while missing the correct strand prediction for three of them. For 7 of the 16 correctly predicted transcripts nocoRNAc predicted very strong transcription signals. The free energy values of the SIDD sites were calculated to be below 4.0 kcal/mol and the confidence values of terminator exceeded 75. According to the authors of TransTermHP such terminators can be considered as high confidence predictions [82]. Three of the loci show an RNAz prediction that is shorter than the database annotation, while it is too long in two other cases. The transcript prediction algorithm of nocoRNAc can compensate for this to some extent and predicts the boundaries of the ncRNAs more precisely than RNAz alone. In addition, nocoRNAc is able to specify the strand of the predicted transcripts. Example loci are shown in figure 4.3. With respect to the tRNAs 57 of 65 loci were detected by RNAz, of which 30 were classified as transcripts by nocoRNAc, including a correct specification of the strand. For another 4 loci the strand was incorrectly predicted. In the Rfam database (version 10.0) 28 *cis*-regulatory elements are annotated for *S. coelicolor*. These elements are part of mRNAs and are not transcribed on their own. As they consist of an evolutionary conserved secondary structure RNAz is able to detect them and nocoRNAc's ncRNA transcript prediction algorithm should be able to distinguish such elements from ncRNAs that are transcribed independently from mRNAs. Of the 28 elements 17 were detected by RNAz and nocoRNAc classifies only one of 17 elements as an ncRNA transcript, which corresponds to a correctness of more than 90%.

Another computational pipeline for the prediction and annotation of ncRNAs in bacteria is SIPHT [93]. SIPHT predicts ncRNAs only in intergenic regions but also takes structure conservation and transcriptional signals, such as Rho-independent transcription terminators and sigma factor binding sites into account. Its general approach is therefore most comparable with that of nocoRNAc. To compare the performance of SIPHT to nocoRNAc's it was applied to the genome of *S. coelicolor* using its web-interface with standard parameters.

For all intergenic regions in the genome of *S. coelicolor* SIPHT predicted 391 ncRNA transcripts. As for the evaluation of nocoRNAc the predictions were compared to annotated ncRNAs. A summary of the results is also provided in table 4.1. Of 86 annotated ncRNA transcripts including tRNAs SIPHT detects 14

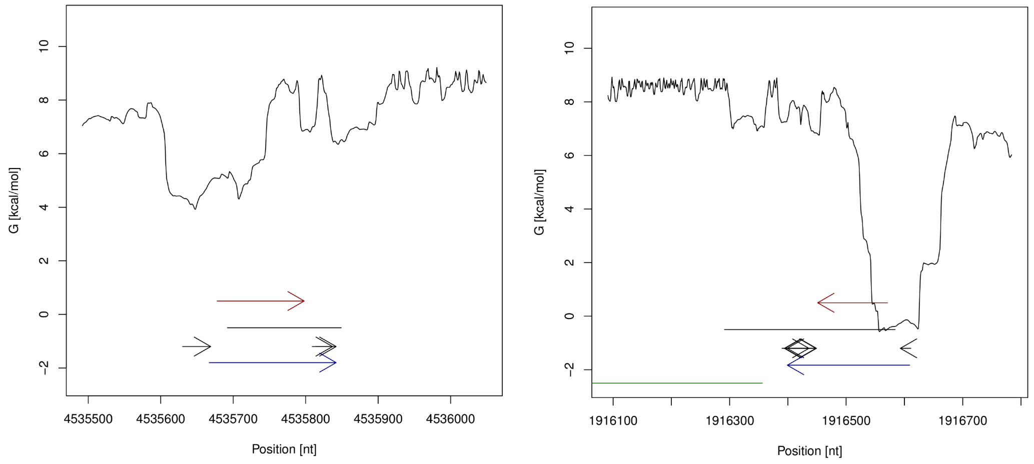


Figure 4.3.: Transcription feature plots of predicted ncRNA transcripts (blue arrows) covering annotated ribosomal RNAs (red arrows). For a detailed legend see figure 4.1. In the first example the RNAz prediction is shorter than the annotated ncRNA (left), while it is much longer in the second example (right). In both cases the prediction of the transcript boundaries was improved by nocoRNAc.

while nocoRNAc is able to detect 46. Especially for tRNAs the sensitivity of SIPHT is quite low as only one of 65 elements is detected. In combination with RNAz nocoRNAc detects 30 tRNAs correctly, which is more than 50%. Furthermore, SIPHT classifies two *cis*-regulatory elements as ncRNA transcripts, which is comparable to nocoRNAc's performance.

4.2. Time series expression analysis of predicted ncRNA transcripts

To verify the expression of ncRNA transcripts predicted for *S. coelicolor*, data from a high resolution time series transcriptome analysis was used, where the effects of phosphate limitation were studied for *S. coelicolor* M145 wild type cultivated in submerged batch fermentations [109]. A custom design microarray was utilized containing 226,576 perfect match oligonucleotide probes that interrogate 8,205 protein-coding regions, 10,834 intergenic regions using a tiling approach, and 3,672 regions antisense to protein-coding genes [13]. During the cultivation phosphate was depleted at 35 h after inoculation. Samples were analyzed for altogether 32 time points covering a 40 h interval from 20 h to 60 h after inoculation.

For the expression profiling of the predicted ncRNA transcripts the array probe sequences were aligned to the *S. coelicolor* genome and the probe interrogation positions were compared to the loci of the predicted ncRNA transcripts. All transcripts that were interrogated by at least four array probes were considered for the subsequent analysis. For each of these transcripts their array probes were grouped in probe sets, which were then added to the CDF descriptor file of the Affymetrix

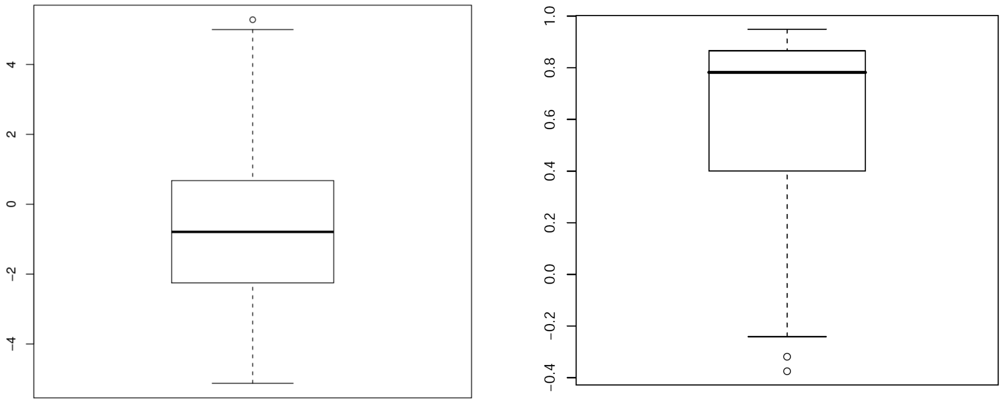


Figure 4.4.: Left: Boxplot of average expression profile differences of predicted asRNAs and their respective protein-coding genes. A negative value indicates a higher expression level of the coding gene. (x = asRNA; y = protein; $d(x, y) = (\sum_{i=1}^{n} x_i - y_i)/n$). Right: Boxplot of expression profile correlations of predicted asRNAs with a variant expression profile and their respective protein-coding gene.

chip. By this, their expression values could be calculated for the different time-points along with those of protein-coding genes. The normalized expression matrix for ncRNAs and coding genes was obtained using RMA as described for the protein-coding dataset [109, 13]. The analysis and visualization of the expression profiles was performed using the transcriptome analysis software Mayday [14].

403 of the 843 predicted ncRNA transcripts fulfilled the criterion of at least four interrogating array probes and their expression was profiled for the 32 time points covering the growth curve of *S. coelicolor* affected by the phosphate limited cultivation conditions. 92 of these elements are located in intergenic regions, whereas the other 311 are located antisense to protein-coding genes. A comparison to annotated ncRNAs showed that 47 of the 92 loci represent putative novel ncRNA transcripts. Altogether 317 of the 403 measured transcripts are considered to be expressed at at least one time-point when using the first quartile of the expression value distribution of the coding elements as a threshold. After application of a variance threshold of 0.025 (regularized variance) 71 of the expressed ncRNA transcripts are considered to show differential expression across the time series.

One important ability of NOCORNAC is the prediction of antisense RNA transcripts. For 235 of the measured antisense RNAs expression could be detected for at least one time-point. For the respective sense-antisense pairs the relation of their absolute expression values was investigated. This was done by calculating the expression value differences of coding genes and their antisense RNAs over the complete expression profile. The result is visualized as a boxplot in figure 4.4 (left). For the majority of sense-antisense pairs the coding gene shows a higher expression than the antisense RNA. However, in about 35% of the cases the expression of the antisense RNA is higher.

47 of the 235 expressed antisense RNAs had a variant expression profile. For these the correlation of the expression profile with that of the respective coding gene was calculated. The resulting correlation value distribution is shown as a boxplot in figure 4.4 (right). It turned out that most sense-antisense pairs show a positive correlation. About 75% of the correlation values are above 0.4 with the median being 0.78. There are no pairs with a strong anticorrelation of expression profiles. Only weak anticorrelation values, which are greater than −0.4, are observed.

To elucidate the different expression patterns that the variant antisense RNA transcripts show during the course of the time series, an unsupervised expression profile clustering was conducted. Profile plots for four expression profile clusters of antisense RNAs together with their respective antisense genes are shown in figure 4.5. The expression of almost all transcripts shows a reaction to the phosphate depletion event at 35h. 24 of the 47 predicted transcripts show a significant downregulation about 1h after the depletion event, which can also be observed for the respective coding genes, of which the majority encodes ribosomal proteins (figure 4.5A).

Four of the predicted antisense RNAs are significantly upregulated immediately after the phosphate depletion (figure 4.5B). Their antisense genes show an even stronger upregulation at the same time. The phosphate binding protein PstS (SCO4142) and the polyphosphate kinase Ppk (SCO4145) are contained in this expression cluster. For these genes it could be shown that they are part of the PhoP regulon [131]. PhoP is a transcription regulator that controls the expression of various genes involved in phosphate uptake and metabolism and shows a strong reaction to phosphate limitation [131].

The genes in the other two expression profile clusters (figure 4.5C/D) are developmental genes that are for example involved in chromosome replication or RNA synthesis. Most of these genes are downregulated as a reaction to phosphate depletion.

The investigation of the 92 predicted ncRNA transcripts that are located in intergenic regions revealed that 82 can be considered as expressed at at least one time point. Of these, 38 do not correspond to an annotated element and are therefore putative novel ncRNAs. Expression profiles of 5 predicted ncRNA transcripts with a differential expression across the time series are shown in figure 4.6. The expression of ncRNA1107_1 is upregulated during the time-course. However, this upregulation does not seem to be related to the phosphate limitation. In contrast, the expression of the two elements ncRNA852_1 and ncRNA2873_1 shows a clear reaction to the phosphate depletion event. They are significantly upregulated about 1h after phosphate is depleted. The expression of ncRNA2823_1 decreases slightly during the time-course without any clear reaction to the depletion event and the profile of ncRNA4158_1 shows a high variance in general but no clear tendency of up- or downregulation.

The predicted ncRNAs have been further assessed with respect to their potential to regulate protein-coding genes by RNA-RNA interactions (see sections 4.3 and 4.4). The putative ncRNA transcript ncRNA2823_1 turned out to be a very good candidate for a non-coding element regulating antibiotic production in *S. coelicolor* (see section 4.4).

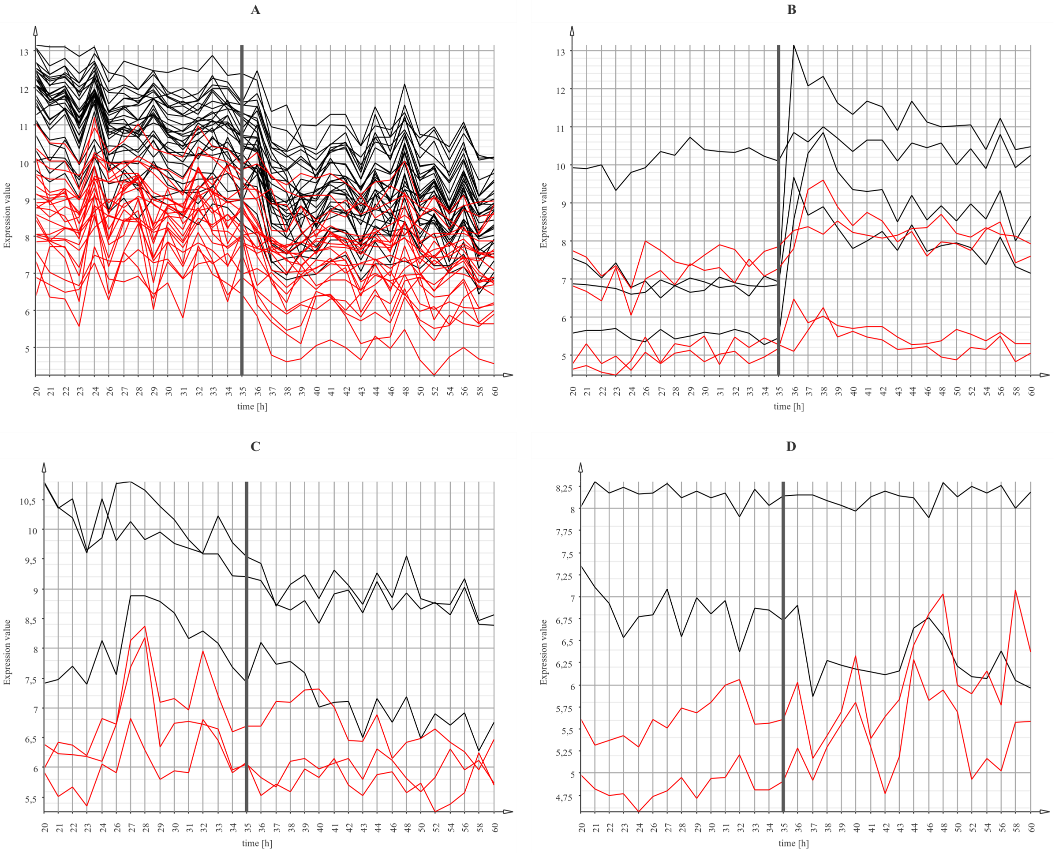


Figure 4.5.: Expression profile plots of 4 clusters of asRNAs (red) and their protein-coding genes (black), which resulted from an unsupervised expression profile clustering. The time-point of phosphate depletion is indicated by a grey vertical line.

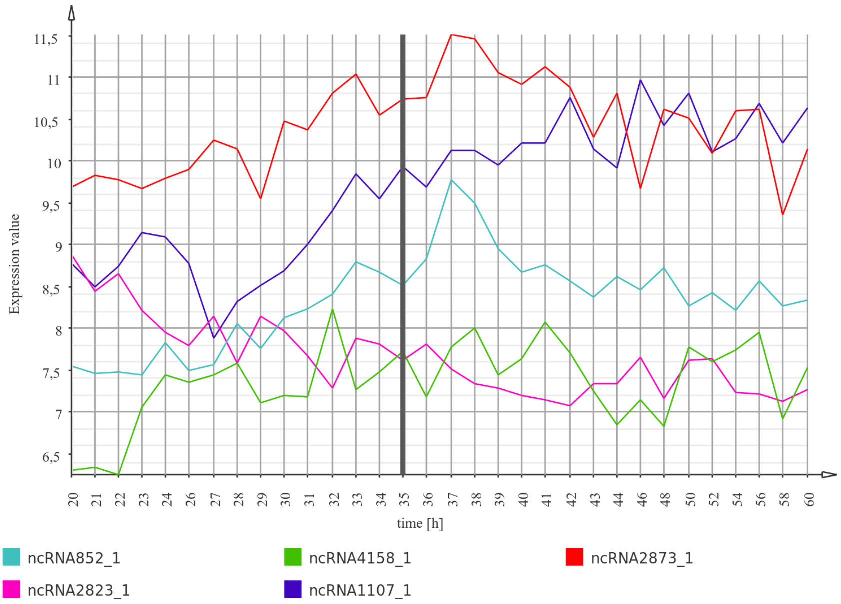


Figure 4.6.: Expression profile plot of predicted ncRNA transcripts that are located in intergenic regions and that show a variant expression profile. The time-point of phosphate depletion is indicated by a grey vertical line.

4.3. Prediction of a network of potential RNA-RNA interactions in Streptomyces coelicolor

As a proof of concept for the methods described in section 3.4 and for a further characterization of the ncRNA loci predicted in the genome of *Streptomyces coelicolor*, their potential to be involved in RNA-RNA interactions with the mRNAs of protein-coding genes was assessed. For this, all predicted ncRNA loci that were shown to be expressed at at least one time-point (expression value ≥ 7.0) during the time series analysis (see section 4.2) and all protein-coding genes that showed a variant expression across the time series (regularized variance ≥ 0.1) were included in this analysis. An RNA-RNA interaction network was predicted for these elements with NOCORNAC by utilizing the RNA-RNA interaction prediction program `IntaRNA` [30]. The results were filtered and evaluated with NOCORNAC's interactive `R` environment (section 3.5). Interaction networks were visualized with `yEd` [180].

Altogether 507 ncRNA loci and 322 protein-coding genes were used as input. For the `IntaRNA` prediction a seed length of 8 was used with 1 mismatch allowed in the seed region. For sequences longer than 100 bp a sliding window approach was applied with a window size of 100 nt (`IntaRNA` parameter `-w`). Furthermore, the flags `-U` and `-P` were set for the application of the `RNAup` model [105] and the `RNAplfold` [18] model during the calculation of the free energy values of the interaction. The resulting interaction predictions were filtered with respect to their free energy values and the length of the interaction site. Only interactions among the best percentile with respect to one of these properties or among the best two percentiles with respect to both properties were kept.

The filtered interaction network contained 491 ncRNA loci, 315 protein-coding genes and altogether 1554 predicted interactions. An overview of the full interaction graph is shown in figure 4.7. 64 of the 1554 high-scoring interactions involve a *cis*-asRNA that was predicted for the respective protein-coding target. All other high-scoring interactions are predicted between *trans*-encoded elements.

The median number of interactions per ncRNA locus is 4. However, it turned out that for some ncRNA loci a very high number of potential interactions was predicted. The subgraph showing the five ncRNA loci with the highest degree is shown in figure 4.8. The median number of predicted interactions per protein-coding gene is 2. An extraordinary high number of interactions was predicted for the three polyketide synthases *cpkABC* (SCO6275, SCO6274, SCO6273), which are part of a type I PKS gene cluster [117]. The cluster has been shown to be involved in the production of a yellow compound [118, 57] and probably in antibiotic activity [57]. However, the functions of the cluster and the produced compound are not fully characterized yet. Time series expression profiles of the three genes in *S. coelicolor* wild type grown under phosphate limited conditions are shown in figure 4.9. They show a strong upregulation at 22 h after inoculation and about 4 h later they are downregulated to their initial expression level. A reaction to the phosphate depletion event, which occurred at 35 h after inoculation, cannot be observed.

The subgraph showing the elements for which a high-scoring interaction with the three PKS genes was predicted is shown in figure 4.10. It can be observed that for some of the ncRNAs interactions with two or even all three genes have been predicted. Furthermore, for each of the 5 ncRNA hubs depicted in figure 4.8 high-scoring interactions were predicted with all three genes. All of these interactions were predicted with *trans*-encoded ncRNAs as no *cis*-encoded asRNAs of SCO6273-SCO6275 were included in the analysis. As described in section 3.4.1 NOCORNAc can calculate RNA-RNA interaction profiles for elements for which several interactions have been predicted in order to assess which of the interactions are the most probable ones. Interaction profile plots for SCO6274 and SCO6275 are shown in figure 4.11. In these plots for each interaction site of the respective mRNA its position is visualized in addition to the free energy and the probability of the interaction. It can be observed that the vast majority of predicted interactions have an extremely low probability. For SCO6274 only one probable interaction has been predicted. The probability of this interaction is almost 1.0. For SCO6275 two probable interactions have been predicted, where one interaction has a probability of about 0.7 and the other one of about 0.3.

This evaluation of the interaction profiles, which are computed by NOCORNAc, shows how they can be applied to determine the most probable interactions for each target within a network of hundreds or thousands of predicted interactions. This allows for choosing only significant interactions for further analyses or experimental validation. It is apparent that the free energy value alone would not be sufficient for this selection as many interactions with similar free energy values have been predicted for SCO6274 and SCO6275. However, it turned out that only three of them have a significant probability. Of course, the interaction with the best free energy

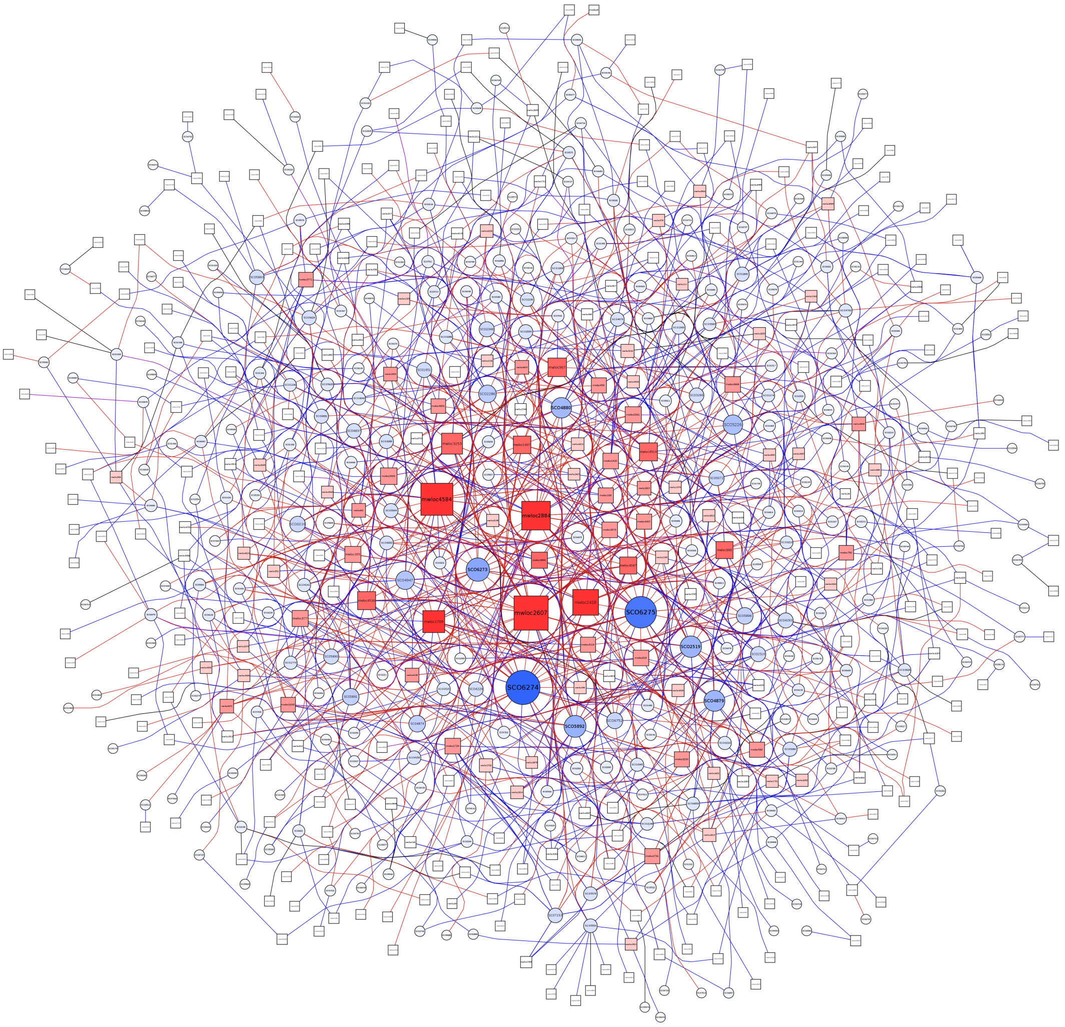


Figure 4.7.: Visualization of a predicted RNA-RNA interaction network in *S. coelicolor* involving – after filtering – 491 predicted ncRNAs, 315 protein-coding genes and 1554 predicted interactions. Protein-coding genes are depicted as blue circles, ncRNAs as red rectangles. The degree of the nodes is denoted by their size and their opacity. Red edges show high-scoring interactions with respect to the free energy value, while blue edges denote interactions with long interaction sites. Interactions shown as purple edges have a good free energy value and a long interaction site. Interactions between a protein-coding gene and its *cis*-encoded asRNA are shown in black. The network was exported as a `gml` file and laid out with `yEd` [180].

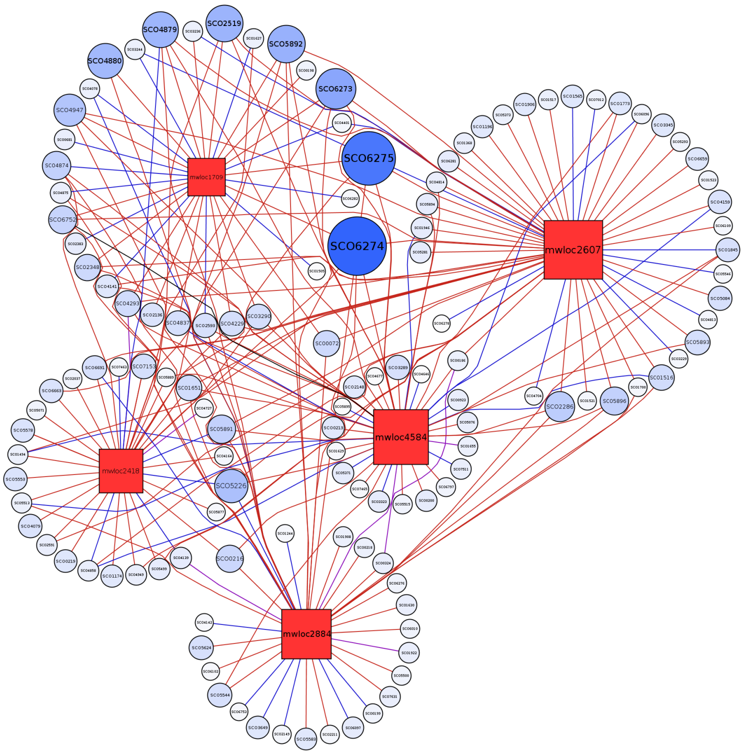


Figure 4.8.: Visualization of the RNA-RNA interaction network involving the five ncRNAs with the highest degree. For a detailed description of the visual elements see figure 4.7.

value is also the most probable one, but that does not mean that this interaction is the only one that has a significant probability as has been observed for SCO6275.

The ncRNA and mRNA hubs that have been identified are promising candidates for elements potentially involved in regulatory processes. To strengthen this hypothesis, thorough assessments of their secondary structures, functional annotations and expression patterns should be the next steps. In the context of *S. coelicolor* particularly the regulation of antibiotics production is of interest. The next section focuses on two predicted ncRNA transcripts that are potentially involved in this process.

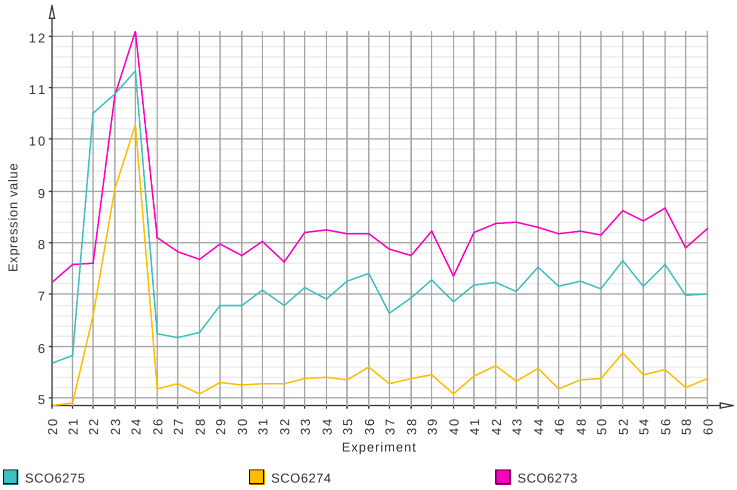


Figure 4.9.: Time series expression profiles of *cpkABC* (SCO6275, SCO6274, SCO6273) in *S. coelicolor* wild type grown under phosphate limited conditions.

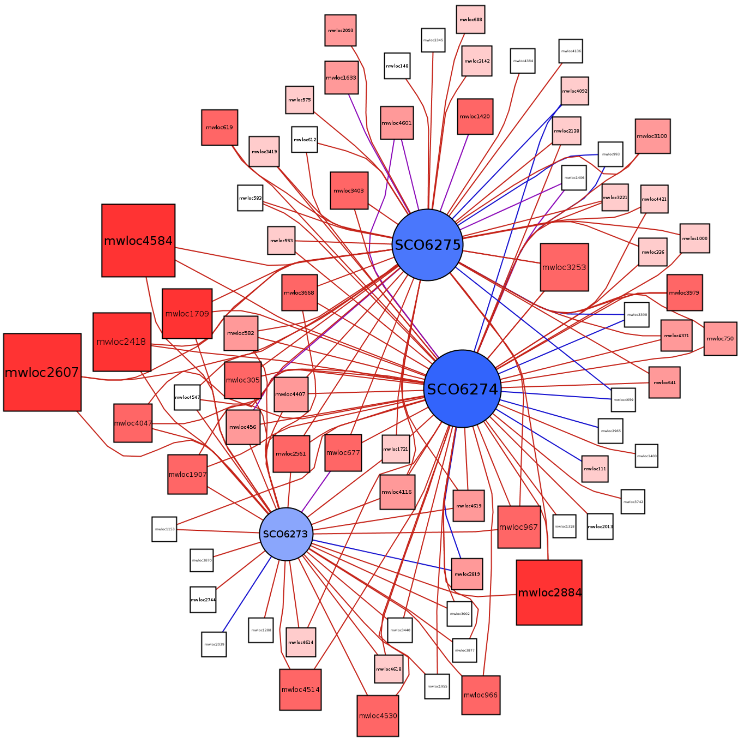


Figure 4.10.: Visualization of the RNA-RNA interaction network involving the three polyketide synthases SCO6273-SCO6275. For a detailed description of the visual elements see figure 4.7.

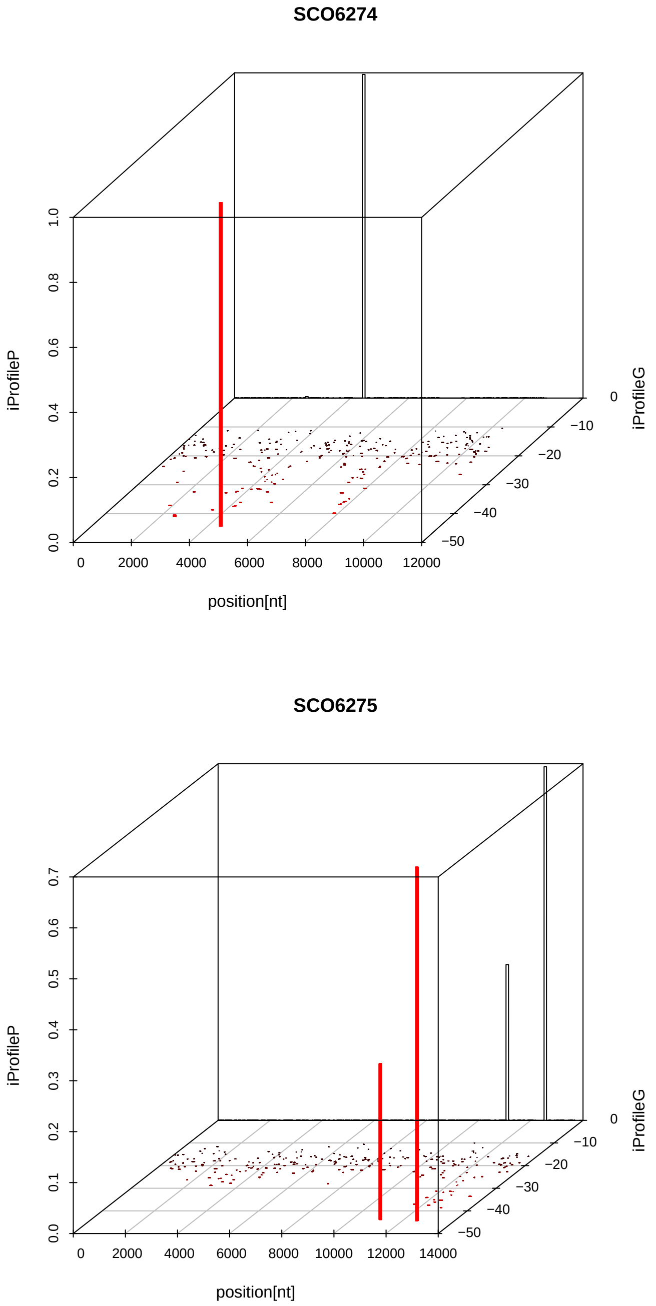


Figure 4.11.: Visualization of RNA-RNA interaction profiles for the two genes SCO6274 (top) and SCO6275 (bottom). The x-axis denotes the position within the gene sequence in base pairs. The y-axis and the z-axis denote iProfileP and iProfileG, respectively. Thereby, for each interaction that was predicted for the respective gene, its position, the free energy value and its probability considering all predicted interactions is shown. It has to be noted that in these plots all predicted interactions are included without restriction to high-scoring interactions. Most of the predicted interactions have only a low probability. For SCO6274 one interaction with very high probability was predicted (almost 1.0), while for SCO6275 two interactions with high probability were predicted (~ 0.3 and ~ 0.7).

4.4. Putative non-coding RNA transcripts potentially involved in the regulation of antibiotic production in Streptomyces coelicolor

In addition to the expression analysis the predicted ncRNA transcripts in the genome of *S. coelicolor* have been further investigated with respect to their potential function as regulators of protein-coding genes. Two of them are extremely promising candidates as regulators of antibiotics production. They are located upstream of the cold shock protein *csp1* (SCO4295) and upstream of *groEL2* (SCO4296), respectively.

It has been shown that the intergenic regions upstream of the corresponding homologs in *Streptomyces hygroscopicus*, if artificially introduced into *S. coelicolor* J1501 using high copy number plasmids, result in an increase in the production of the antibiotic compounds actinorhodin (ACT) and undecylprodigiosin (RED) [100]. Furthermore, the authors demonstrated that the introduction of the fragments has a direct or indirect influence on the expression of the actinorhodin pathway-specific regulator *actII-ORF4* (SCO5085), which is increased in strains carrying the plasmid. Thus, ncRNAs that might be contained in these regions are potentially targeting the respective antibiotic pathway regulators *actII-ORF4* (SCO5085) and *redD/Z* (SCO5877/SCO5881) or their regulators.

Several genes have been identified that potentially act as pleiotropic regulators of antibiotic production in *S. coelicolor* [78]. As a candidate for a global downregulator *wblA* (SCO3579) has been suggested. A probable global upregulator is the putative sigma factor SCO5147. As another antibiotic downregulator that functions independently from *wblA* the putative TetR family transcriptional regulator SCO1712 has been identified [88]. Overexpression of SCO1712 leads to a repression of the antibiotic pathway-specific regulators *actII-ORF4* (SCO5085) for ACT, *redD/Z* (SCO5877/SCO5881) for RED and *cdaR* for CDA (calcium-dependent antibiotic) and thus to a significant reduction of antibiotic production.

The predicted ncRNAs that are located in the respective intergenic regions were carefully analysed for transcriptional features. In addition, their expression was profiled for the complete time series of *S. coelicolor* grown under phosphate limited conditions, and protein-coding genes were identified that show similar expression patterns.

To specify the potential regulatory function of the two putative ncRNA transcripts, RNA-RNA interaction predictions were performed between the ncRNAs and protein-coding genes with functional annotations including the known antibiotic regulators listed above.

To increase the confidence of high scoring interaction candidates by compensating the effects of transcript lengths and base pair frequencies z-scores and p-values were calculated for predicted interactions.

4.4.1. Transcriptional features of predicted csp-ncRNA

The structurally conserved region upstream of *csp1* as predicted by RNAz has a significant P-value of 0.94. nocoRNAc was used to analyse this region for tran-

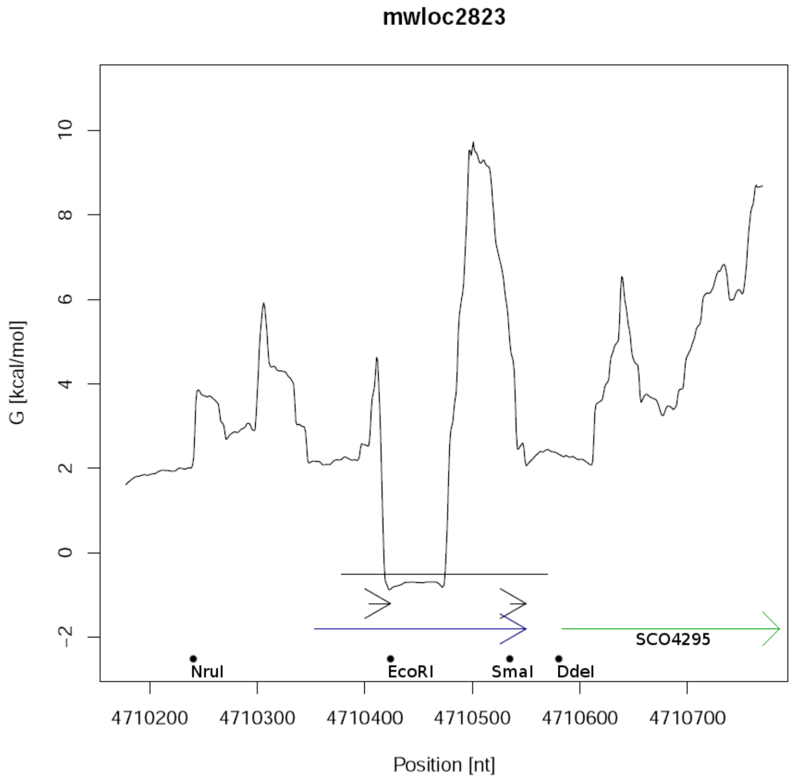


Figure 4.12.: Transcription feature plot of the csp-ncRNA (ncRNA2823_1). The region predicted by RNAz is drawn as a black line. The SIDD profile is shown as a black graph. Small black arrows represent predicted terminator signals. The transcript predicted by NOCORNAc is drawn as a blue arrow. The green arrow denotes the position of the *csp1* CDS (SCO4295). Black dots show the restriction sites in the original *S. hygroscopicus* fragment mapped to the genome of *S. coelicolor*.

scriptional features. A SIDD site with a free energy value of about 2.0 kcal/mol was found upstream of the region and thus chosen as the predicted transcription start for this ncRNA. Furthermore, a predicted Rho-independent terminator with a confidence value of 51 was found within the region. The resulting predicted ncRNA transcript (csp-ncRNA) is of 198 nt length and it is located upstream of the cold shock protein *csp1* (SCO4295). It ends 34 nt upstream of the translation start site of *csp1*. A plot showing all predicted transcriptional features for this region is shown in figure 4.12.

4.4.2. Effects of different DNA fragments of S. hygroscopicus on antibiotics production in S. coelicolor

Several different DNA fragments originating from the intergenic region upstream of *csp1* in *S. hygroscopicus* have been introduced into *S. coelicolor* J1501 to determine the exact region necessary to affect antibiotic production [100]. Different restriction sites were used to generate several different fragments.

Table 4.2.: Effects of fragment *Eco*RI-*Dde*I on antibiotic production in *S. coelicolor* J1501 [100].

Antibiotic	Control	Carrying Fragment
Actinorhodin	100%	140%
Undecylprodigiosin	100%	139%

The antibiotic production of the strain carrying the fragment is stated relatively to the unmodified strain J1501. The fragment causes an increase in the production of Actinorhodin and Undecylprodigiosin by a factor of about 1.4.

The size of the genomic region in *S. hygroscopicus* from which fragments have been introduced has a length of 1899 bp. The chromosomal coordinates of the fragments are defined by the respective restriction sites of the restriction enzymes. To map the relevant fragments to the genome of *S. coelicolor* the nucleotide sequence of the original genomic region was first digested *in silico* to get the location of the fragments relative to the genomic region of *S. hygroscopicus*. To determine the coordinates relative to the genome of *S. coelicolor* a BLAST search has been conducted with the *S. hygroscopicus* region as query and the genomic sequence of *S. coelicolor* as target. The original fragment positions were then mapped to the *S. coelicolor* genome using the resulting alignment.

The restriction sites that are relevant for the relative position of the csp-ncRNA are shown in figure 4.12. The two fragments, *Nru*I-*Dde*I and *Eco*RI-*Dde*I, led to an increased antibiotics production. Fragments that end at *Sma*I did not influence antibiotics production.

The sequence similarity between *S. coelicolor* and *S. hygroscopicus* is quite high for the *Eco*RI-*Dde*I fragment (about 89%), whereas it is significantly lower for the region between *Nru*I and *Eco*RI (about 51%).

The effects of the fragment *Eco*RI-*Dde*I on antibiotic production in *S. coelicolor* J1501 are shown in table 4.2. The production of ACT and RED is increased by a factor of about 1.4.

4.4.3. Expression of the csp-ncRNA during the phosphate limited time series

An expression profile of the csp-ncRNA is shown in figure 4.13. 20h after inoculation the expression is at a higher value (about 9.0) and it decreases until it reaches its minimum at 42h after inoculation (about 7.0). Among the protein-coding genes with a similar expression profile is the transcriptional regulator *glnR* (SCO4159) and the two genes *cobN* (SCO1849) and *cobO* (SCO1851), which are involved in cobalamin biosynthesis.

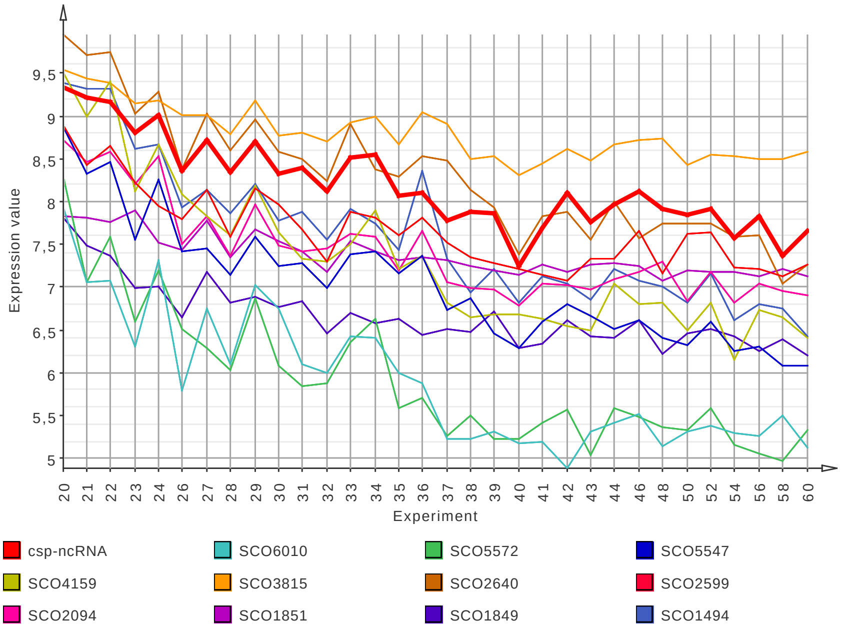


Figure 4.13.: Time series expression profile of csp-ncRNA (thick red profile) together with the expression profiles of protein-coding genes showing a similar expression pattern (Pearson Correlation distance ≤ 0.1).

4.4.4. Potential RNA-RNA interaction targets of the csp-ncRNA

To identify possible targets of the csp-ncRNA RNA-RNA interaction predictions have been conducted using the `intarna` function integrated in NOCORNAC's R environment (see section 3.4.3).

In the target prediction for the csp-ncRNA all protein-coding genes with an annotated function were considered. The complete sequence of the ncRNA and the complete CDS of the genes were used as input.

z-score and p-value calculations have been conducted for the 10 highest scoring predicted RNA-RNA interaction partners of the csp-ncRNA that are annotated to have a regulatory function. The results are listed in table 4.3. Furthermore, interaction prediction statistics were computed for the antibiotic regulators *actII-ORF4* (SCO5085), *redD/Z* (SCO5877/SCO5881), *wblA* and SCO5147. For each gene the free energy value of the predicted interaction with the csp-ncRNA is given as well as a z-score and p-values calculated for this prediction.

No strong interaction was predicted for the pathway-specific regulator genes of ACT and RED or the global antibiotic regulators *wblA* or SCO5147. However, a very good interaction was predicted for the global antibiotic regulator TetR (SCO1712). The confidence of the low free energy value of about −17.8 kcal/mol is strengthened by a low z-score and a very small p-value. A summary of the details of this predicted interaction is provided in table 4.4.

In addition, there are some more (putative) regulatory genes, for which a strong interaction was predicted. SCO1587 is annotated as a GntR family protein. Detailed

Table 4.3.: Regulatory genes potentially targeted by the csp-ncRNA via RNA-RNA interaction.

SCO ID	annotation	length	energy	z-score	p-value
SCO1587	GntR	1151	-20.4517	-3.79	3.06e-13
SCO5828	Response regulator	628	-19.4142	-6.71	8.77e-18
SCO6696	Other regulation	2448	-19.3817	-3.76	3.47e-13
SCO5460	Other regulation (AbaA)	608	-18.3866	-6.98	4.21e-18
SCO1897	DeoR	691	-17.8802	-4.84	3.69e-15
SCO1712	**TetR**	**461**	**-17.7836**	**-3.81**	**2.79e-13**
SCO3664	Other regulation	743	-17.7031	-4.00	1.14e-13
SCO3556	AraC	1239	-17.4250	-3.92	1.64e-13
SCO0194	sigma factor	606	-17.2942	-2.77	7.31e-11
SCO7767	Other regulation	825	-17.1686	-4.14	6.19e-14
SCO5085	actII-ORF4	767	-8.79622	0.98	0.99
SCO5877	redD	1052	-10.9028	0.11	0.69
SCO5881	redZ	653	-10.2637	-0.79	0.0010
SCO3579	wblA	338	-10.4563	-0.10	0.31
SCO5147	putative sigma factor	671	-9.47578	0.57	0.99

For each potential target gene the locus tag, gene name or functional annotation, CDS length and details about the predicted interaction with the csp-ncRNA are provided. Interaction details include the free energy value of the interaction predicted by `IntaRNA` and the respective z-score and p-value as calculated by NOCORNAc.

information about the specific function of this gene is not available. However, the DasR regulator, which belongs to a subclass of this family, has been shown to be a global regulator of primary metabolism and development in *S. coelicolor* [128]. It has also been shown to influence antibiotic production, but there was no strong interaction predicted between the csp-ncRNA and the DasR mRNA.

SCO1897 is a protein of the DeoR family. DeoR-like transcription repressors occur in diverse bacteria as regulators of sugar and nucleoside metabolic systems (Pfam). Therefore, it is very unlikely that a specific regulation of the antibiotic production involves this gene.

Expression profiles of all genes listed in table 4.3 are presented in figure 4.14. Almost all genes, for which a good interaction was predicted, do not show a significant differential expression across the time series, although most of them are clearly expressed. This is also true for *tetR* (SCO1712). Only the putative sigma factor SCO0194 shows a significantly higher expression at the beginning of the time-course than at later time-points and interestingly the profile is fairly similar to the expression profile of the csp-ncRNA.

These results suggest that the csp-ncRNA might potentially regulate TetR via RNA-RNA interaction, as TetR is the only protein, for which a high interaction probability was predicted and for which it has been shown, that it clearly acts as a global antibiotic downregulator. If the csp-ncRNA represses TetR, this could explain why a DNA fragment that contains this predicted ncRNA is able to increase antibiotic production significantly when introduced into *S. coelicolor*.

To investigate this possible interaction further the secondary structure of the csp-ncRNA was predicted and the potential interaction site with the TetR mRNA has been determined. The RNAfold web server was used to predict and visualize

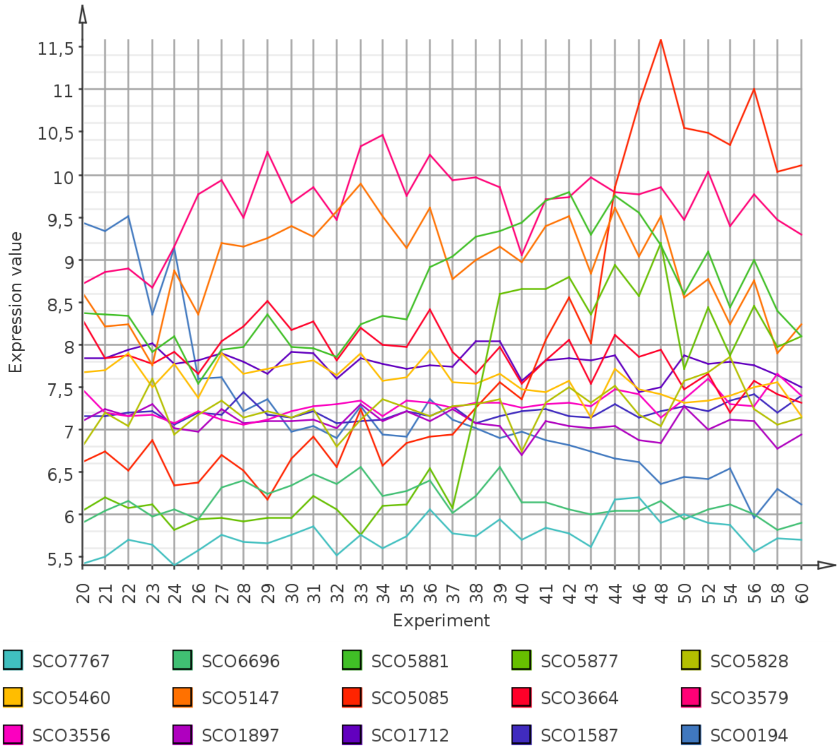


Figure 4.14.: Time series expression profiles of regulatory genes predicted to be potentially targeted by the csp-ncRNA in *S. coelicolor* wild type grown under phosphate limited conditions.

Table 4.4.: Details on the predicted RNA-RNA interaction between the csp-ncRNA and the TetR mRNA.

ncRNA length	618 bp
target length	198 bp
position of interaction site ncRNA	120-146
position of interaction site target	234-265
ED ncRNA	13.3 kcal/mol
ED target	15.2 kcal/mol
hybridization energy	-46.2 kcal/mol
interaction energy	-17.8 kcal/mol
z-score	-3.81
p-value	2.79e-13

The *ED* values denote the energy needed to make the interaction site accessible in the ncRNA and the target mRNA.

secondary structures (standard parameters) [62]. The position of the interaction site was parsed from the `IntaRNA` output. The visual indication of the site in the secondary structure plots was done manually. The result is shown in figure 4.15. In the MFE structure the site contains two unpaired regions, one at its beginning and one at its end, which consist of about 7 nt each. This can probably be sufficient for the initiation of an interaction. The predicted site also spans a stem in the MFE structure. However, the base pair probabilities of this stem are relatively low. Thus, this region might still be involved in the duplex formation with a potential interaction target. Altogether the predicted interaction site consists of 27 nt in the csp-ncRNA and 32 nt in the TetR mRNA, where it is located in the middle of the sequence.

The position of the interaction site in the csp-ncRNA is 52 nt upstream of the 3' end. Therefore, the site would also be contained in the ncRNA if it starts with its alternative SIDD site located at the start of the *Eco*RI-*Dde*I fragment. To confirm this an additional RNA-RNA interaction prediction was performed between this shorter version of the ncRNA (EcoRI-csp-ncRNA) and the TetR mRNA. The result is shown in figure 4.16. The predicted interaction site is 4 nt shorter but apart from this the position of the site is the same. It still contains two unpaired regions at its start and end. Thus, also the topology of the site has not changed significantly. The base pairing probabilities of the stem that is spanned by the site in this MFE are much higher, though.

Interestingly, the position of the interaction site in the csp-ncRNA is almost the same for the predicted interactions with the mRNAs of other regulatory genes listed in table 4.3, such as SCO1587 and SCO1897. However, the base pairings of these interactions are more different.

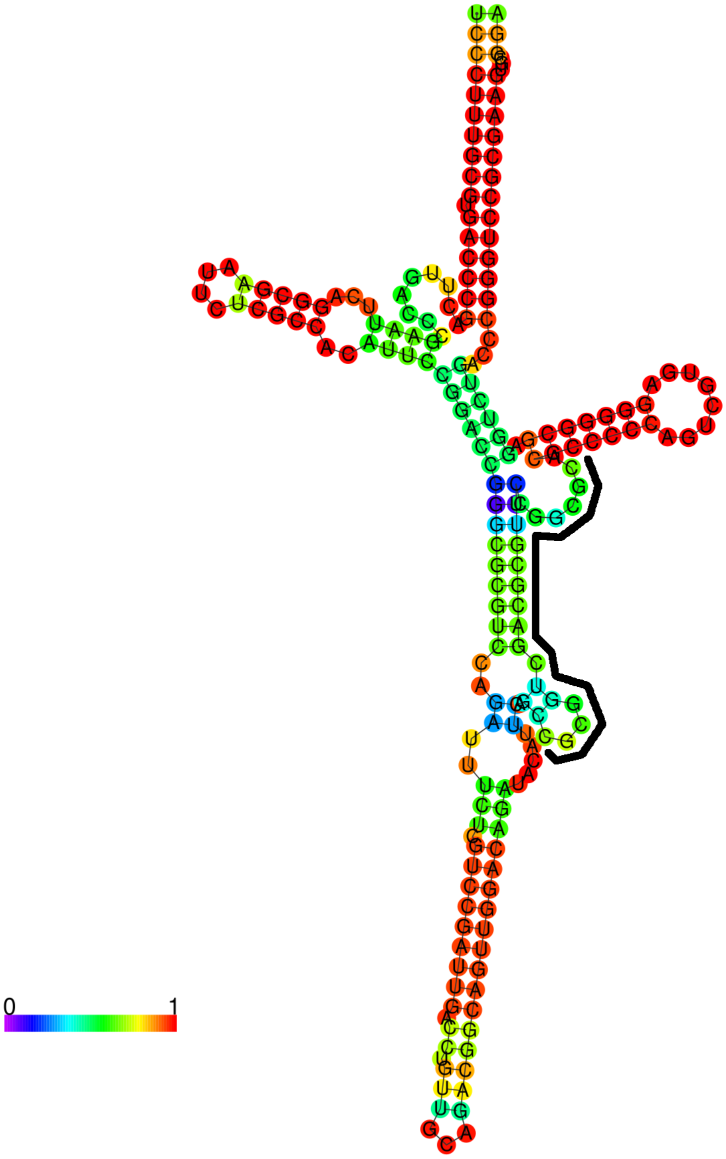


Figure 4.15.: Predicted secondary structure of the csp-ncRNA with indicated predicted interaction site with the *tetR* mRNA (black line). The color-code represents base pairing probabilities.

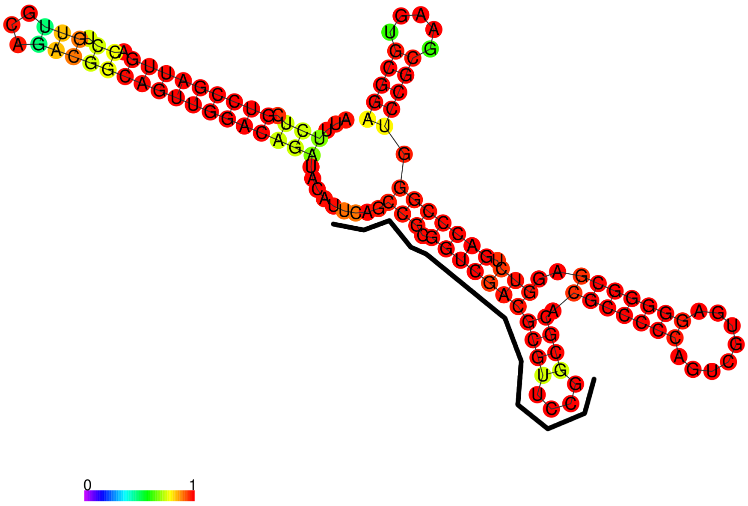


Figure 4.16.: Predicted secondary structure of the EcoRI-csp-ncRNA with indicated predicted interaction site with the *tetR* mRNA (black line). The color-code represents base pairing probabilities.

4.4.5. Predicted ncRNA upstream of groEL2 may also influence antibiotic production

Upstream of SCO4296 (*groEL2*) another ncRNA transcript was predicted (groEL-ncRNA). A transcription feature plot of the respective genomic region is shown in figure 4.17. The groEL-ncRNA overlaps the CDS of SCO4296. It has been shown that the intergenic region upstream of *groEL2* in *Streptomyces hygroscopicus*, if artificially introduced into *S. coelicolor* J1501, results in a similar increase in antibiotic production as the introduction of the *Eco*RI-*Dde*I fragment [100]. The introduction of a fragment containing both loci results in an even stronger increase. It has to be mentioned that this larger fragment also contains the complete CDS of *csp1* (SCO4295). The effects of the different fragments on antibiotic production in *S. coelicolor* are summarized in table 4.5.

An expression profile of the groEL-ncRNA is shown in figure 4.18.

RNA-RNA interaction prediction between the groEL-ncRNA and the TetR (SCO1712) mRNA resulted in a medium free energy value of -11.46 kcal/mol, with a significant z-score of -2.96 and a very small p-value of $2.377e - 11$.

Table 4.5.: Effects of fragment *Eco*RI-*Dde*I (A), fragment containing the groEL-ncRNA (B) and fragment containing both loci (C) on antibiotic production in *S. coelicolor* J1501 [100].

Antibiotic	**Control**	**Carrying Fragment**		
		A	**B**	**C**
Actinorhodin	100%	140%	157%	255%
Undecylprodigiosin	100%	139%	154%	294%

The antibiotic production of the strains carrying the respective fragments is stated relatively to the unmodified strain J1501.

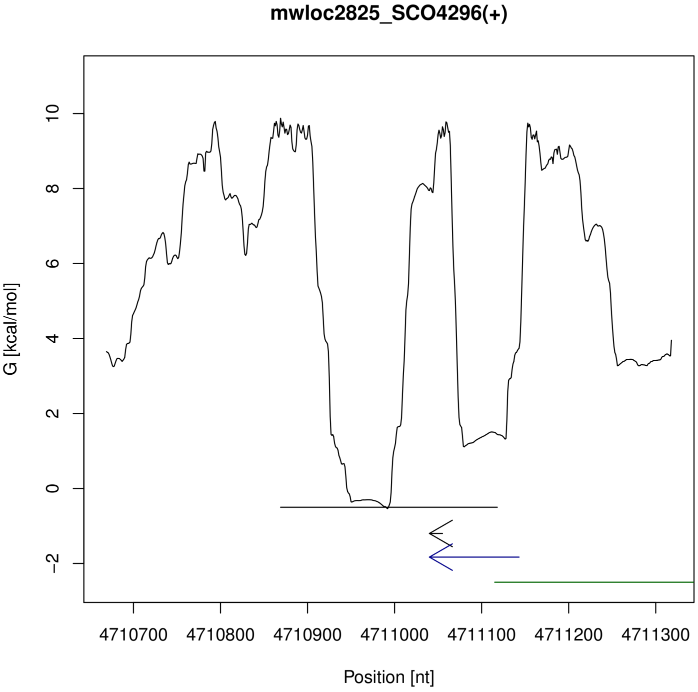


Figure 4.17.: Transcription feature plot of groEL-ncRNA (ncRNA2825_1). The green arrow denotes the position of the *groEL2* CDS (SCO4296). For a detailed legend see figure 4.12.

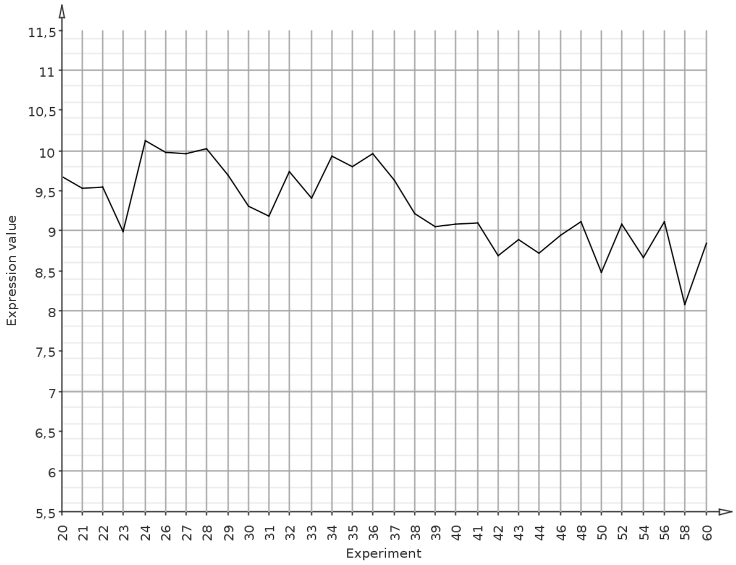


Figure 4.18.: Time series expression profile of the predicted groEL-ncRNA in *S. coelicolor* wild type grown under phosphate limited conditions.

4.5. Differential gene expression in a S. coelicolor GlnK mutant

In *Streptomyces coelicolor* the protein GlnK is an important nitrogen sensor and regulator of genes involved in nitrogen metabolism. Furthermore, it also has a regulatory influence on morphological differentiation and production of secondary metabolites. To identify genes that show a direct or indirect reaction to the depletion of nitrogen in the cultivation medium and that are regulated by GlnK, two transcriptome time series experiments have been conducted and analysed. In the previous sections data from time series experiments have been investigated, where the *S. coelicolor* wild type was grown under phosphate limited conditions. In this section, one time series is presented, where *S. coelicolor* wild type (M145) was grown under glutamate limited conditions and a second time series, where a *glnK* inactivation mutant (SC*glnK*-3 [69]) was cultivated under the same conditions. A comparative analysis of the two time series with respect to the expression of protein-coding genes has been published in [164]. In this section these results are summarized.

For the wild type strain altogether 30 samples were analysed covering a time interval from 20h to 58h after inoculation. The resolution was one hour from 24 to 32h and from 40 to 42h, and half an hour from 32 to 40h. After 40h samples were analysed for 44, 46, 54, and 58h after inoculation. For the SC*glnK*-3 mutant 16 samples were considered in four hour resolution from 21 to 33h, in one hour resolution from 33 to 40h, and in two hour resolution from 40 to 50h after inoculation. The glutamate depletion event was at 34.5h in the wild type and at 34h in the mutant strain.

The time series data analysis tool Tiala [77], which is integrated in the transcriptome analysis software Mayday [14], was used for the comparative analysis and visualization of the two transcriptomic time series. Mayday was used for the expression profile clustering and for visualization. A self-written script was used to perform an initial screening of all protein-coding genes to determine if they are expressed in only one or both time series and if the expression levels differ. For protein-coding genes showing a variant expression profile in at least one of the time series the script evaluates the correlation of the two expression profiles to decide if a gene shows a similar expression pattern in both time series or not. The results of this pre-screening were evaluated manually using the comparative expression profile visualization methods integrated in Tiala.

4.5.1. Unsupervised expression profile clustering of variant genes in the wild type strain

As a first aspect the expression patterns that were exhibited during the time-course of the wild type cultivation and the reaction of genes to the glutamate depletion event were investigated. For this, an expression profile clustering (QT-Clustering) of all 651 genes with a variant expression profile across the time series (regularized variance $>$ 0.05) was performed. Most of these 651 genes are either down- or upregulated around the time of glutamate depletion. 313 genes are downregulated around the time of

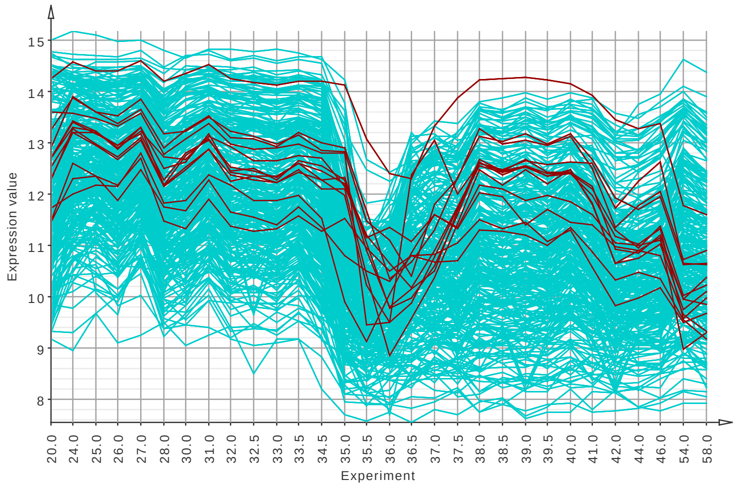


Figure 4.19.: Expression profiles of genes that show a downregulation around the time of glutamate depletion in the wild type strain. Genes downregulated at 35h after inoculation are shown in blue. Genes downregulated at 35.5h after inoculation are highlighted in red.

glutamate depletion. The vast majority of these genes starts at a high expression level and shows a strong downregulation at 35h after inoculation (figure 4.19). It consists of many genes involved in the TCA cycle or glycolysis as well as amino acid metabolism. The glutamate transporter encoding genes *gluABCD* (SCO5774-SCO5777), the *nar2* operon (SCO0216-SCO0219) and the ATP synthase gene cluster (SCO5366-SCO5374) are also contained in this group as well as a large number of ribosomal protein encoding genes. A small group of genes shows nearly the same expression pattern, but the downregulation at the glutamate depletion event occurs half an hour later at 35.5h (highlighted in figure 4.19). The group contains, for example, genes involved in sulfate assimilation (*cysH*, *cysCDN*; SCO6097-SCO6100).

307 genes are upregulated around the time of glutamate depletion or shortly afterwards. One cluster of genes shows even an earlier reaction. These genes are upregulated at 24h after inoculation and are downregulated about 3h later (figure 4.20A). At 35h they show a second upregulation for about 1 to 1.5h. Contained in this group is for example the putative TetR family regulator SCO3207. A second group of genes shows an upregulation at 35h, which lasts for only about half an hour (figure 4.20B). Among these is for example the sigma factor SCO7314. Another group of genes is also upregulated at 35h but for up to 1.5h (figure 4.20C). In this group many genes involved in fatty acid metabolism can be found as well as various transporters (e.g. sugar transporters) and the ectoine biosynthesis genes. The largest group of genes shows an upregulation at 35h which lasts up to 6h (figure 4.20D). Among these 107 genes there are again many transporters and genes involved in amino acid metabolism. In addition, some genes of the TCA cycle and the NADH dehydrogenase cluster (*nuo* genes; SCO4562-SCO4575) are contained in this group. A small group of genes is upregulated at 35.5h for up to 1h (figure 4.20E). It contains for example the phenylacetic acid degradation protein encoding genes SCO7469-SCO7474

(*paa* genes). Another small group of genes is upregulated at 36h and downregulated about 1h later (figure 4.20F). It contains several transporters and the acetoacetyl-CoA synthetase encoding gene *acsA* (SCO1393). Two further groups of genes show a later upregulation that does not seem to be directly connected to the glutamate depletion event. The first group is upregulated at 38h for about 3h (figure 4.20G). It contains the sigma factor *sigU* (SCO2954) and several membrane protein encoding genes. The second group, which shows an upregulation at 44h, that lasts until the end of the time series (figure 4.20H), mainly consists of the actinorhodin biosynthesis gene cluster.

4.5.2. Expression profile analysis of genes related to nitrogen metabolism in the wild type strain

A specific analysis of genes known to be related to nitrogen metabolism was performed in the wild type strain under glutamate limited conditions.

An expression-based clustering of these 41 genes revealed that 20 show a change in expression level at the time of glutamate depletion (35h). Six of these genes show an upregulation. *narB* (SCO7374) and *gltB* (SCO2026) are among them, although they show only a slight upregulation. The most striking upregulation is exhibited by *glnA4* (SCO1613), which, after a strong upregulation at 35h, is downregulated after 40h. 14 genes are downregulated at the time of glutamate depletion. This group includes among others the operons *nar2* (SCO0216-SCO0219) and *nar3* (SCO4947-SCO4950), the regulator *glnR* (SCO4159) and the gene of glutamine synthetase I *glnA* (SCO2198). The strongest downregulation is shown by the genes of the *nar2* operon. Interestingly, some of the downregulated genes are upregulated soon after glutamate depletion. This is especially the case for the *nar3* genes, which show a strong upregulation immediately after downregulation. Other genes show a similar upregulation, e.g. *glnR* and *glnA*.

Expression profiles of these genes are shown in figures 4.21 and 4.22.

4.5.3. Genes regulated in the wild type strain and not differentially regulated in the SC*glnK*-3 mutant

101 genes were identified that are upregulated in the wild type strain around the time of glutamate depletion and that are almost not differentially regulated in the SC*glnK*-3 mutant. I.e., their expression remains at a constant level or they are only slightly upregulated in the SC*glnK*-3 mutant but their upregulation in the wild type strain is significantly stronger.

Among these genes is for instance the NADH dehydrogenase cluster (SCO4562-SCO4575). In the SC*glnK*-3 mutant this group shows an increase in expression from the beginning of the time series until about 40h after inoculation. Then a slight downregulation is observed until the end of the time series. In the wild type strain these genes remain at a constant expression level at the beginning of the time-course until the time of glutamate depletion when they are strongly upregulated by about

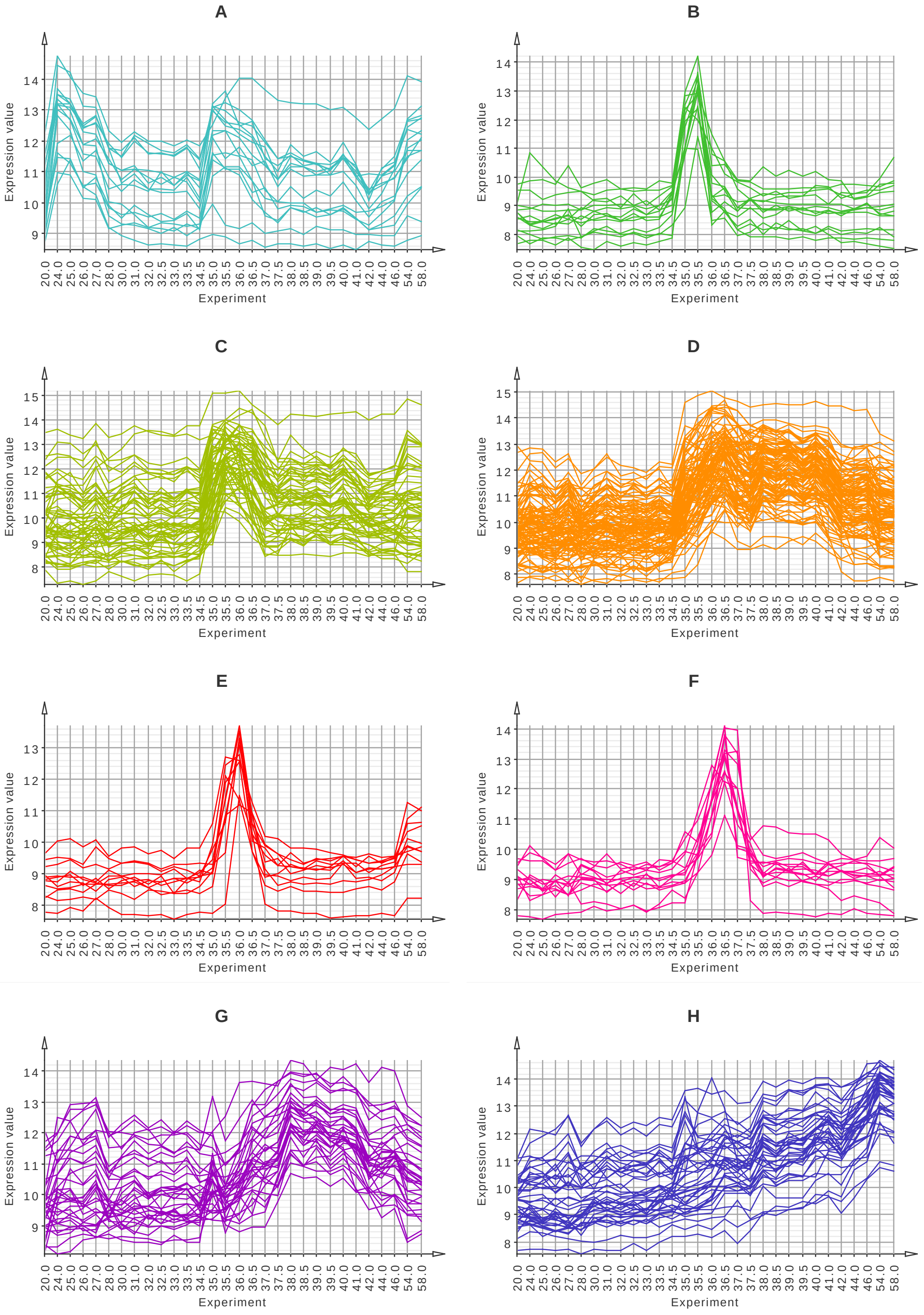


Figure 4.20.: Expression profiles of different gene expression clusters that show an upregulation around or after the time of glutamate depletion in the wild type strain.

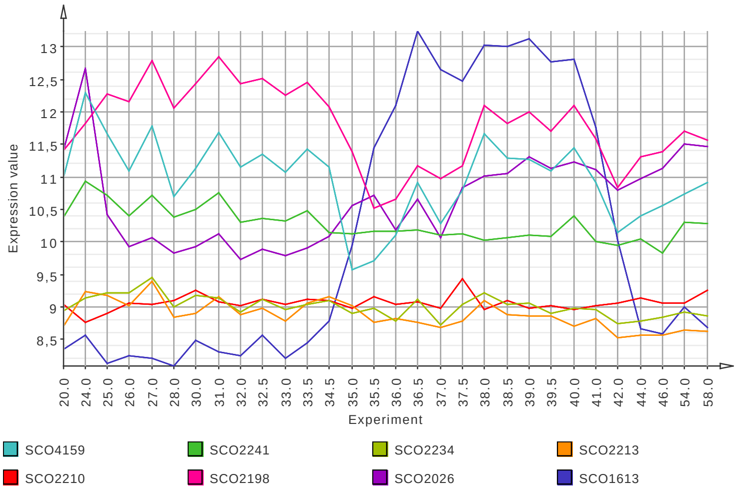


Figure 4.21.: Expression profiles of genes involved in nitrogen metabolism in the wild type strain.

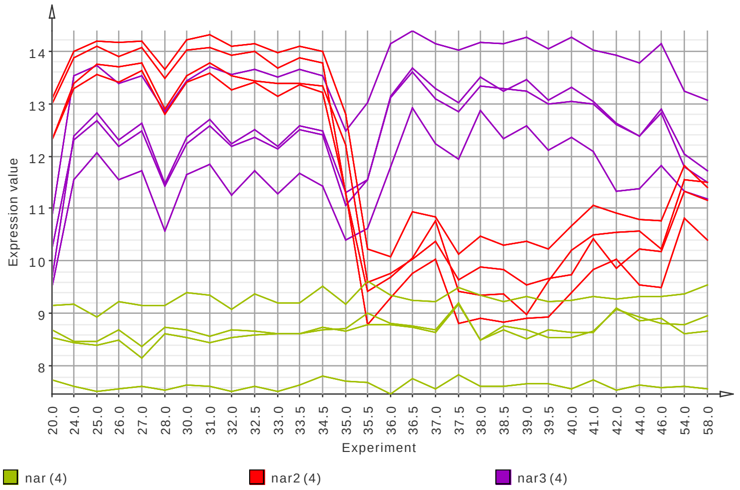


Figure 4.22.: Expression profiles of the *nar*, *nar2* and *nar3* operons in the wild type strain.

2 fold changes on logarithmic scale. This strong upregulation is not observed in the SC*glnK*-3 mutant.

A significant differential expression is also observed for the gas vesicle synthesis genes *gvpOAF* (SCO6499-SCO6501). In the wild type strain *gvpA* and *gvpF* show a strong upregulation by almost 3 fold changes at the time of glutamate depletion. About 3 hours later they are downregulated to their initial expression level. In the SC*glnK*-3 mutant the expression of the two genes remains almost constant throughout the time series. Interestingly, *gvpO* is clearly upregulated in both time series at the time of glutamate depletion, but the increase in expression strength is significantly stronger in the wild type strain.

Another group of differentially expressed genes are the *ramCSABR* genes (SCO6681-6685). At the time of glutamate depletion the *ramCSA* genes are significantly upregulated in the wild type strain but they show a constant expression level throughout the time series in the SC*glnK*-3 mutant. For *ramB* no reaction to glutamate depletion is observed in both time series. *ramR* is upregulated after glutamate depletion in the SC*glnK*-3 mutant and the wild type strain but in the wild type strain the upregulation is stronger.

4.5.4. Genes regulated in the SC*glnK*-3 mutant and not or only slightly differentially regulated in the wild type strain

42 genes were identified that are clearly upregulated in the SC*glnK*-3 mutant around the time of glutamate depletion and that show a significantly different expression in the wild type strain. Among these are for example the *ragABKR* genes (SCO4072-SCO4075). The *ragABK* genes show a very strong upregulation of 2 to 3 fold changes immediately after glutamate depletion in the SC*glnK*-3 mutant. These genes are also upregulated after this event in the wild type strain but change in expression strength is significantly weaker (about 1 fold change). The regulator gene *ragR* is only slightly upregulated in the SC*glnK*-3 mutant after glutamate depletion while its expression level is almost constant throughout the time-course in the wild type strain.

4.5.5. Differentially regulated genes of the TCA cycle and Glycolysis

An analysis of the genes involved in the TCA cycle revealed that most of them are not differentially regulated when comparing the SC*glnK*-3 mutant to the wild type strain. An exception are the two succinate dehydrogenase genes SCO5106 and SCO5107. They show a strong upregulation in the wild type strain around the time of glutamate depletion (about 2.5 fold changes). Then they remain at a very high expression level until 40h after inoculation. In the SC*glnK*-3 mutant these two genes show only a weak upregulation around the time of glutamate depletion (below one fold change, see figure 4.23). Interestingly, the expression profile of *glnA4* (SCO1613), which is very similar to the profiles of SCO5106 and SCO5107 in the wild type strain, shows a different behaviour in the SC*glnK*-3 mutant.

Most of the genes involved in glycolysis are also not differentially regulated between the SC*glnK*-3 mutant and the wild type strain. However, SCO5983 and

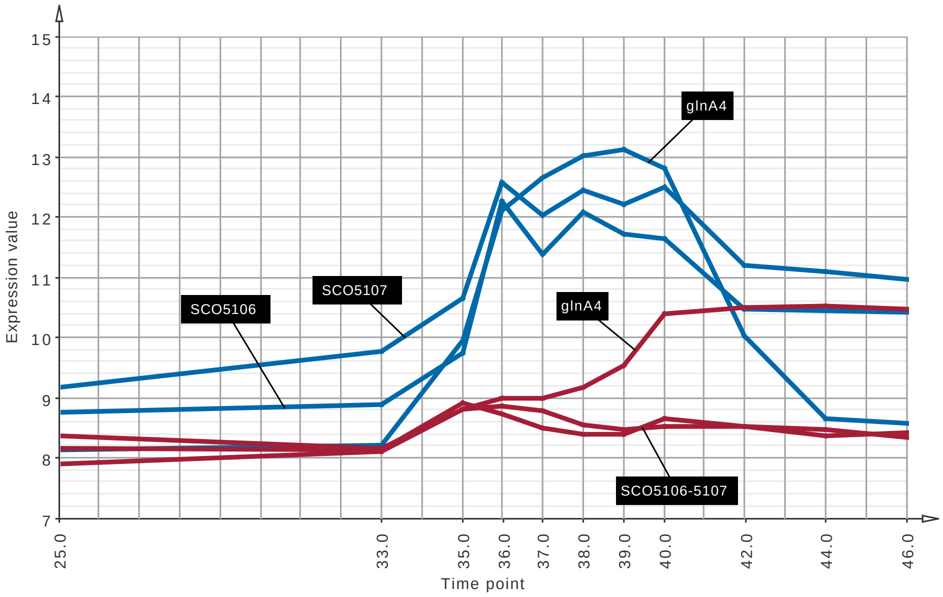


Figure 4.23.: Expression profiles of the two succinate dehydrogenase genes SCO5106 and SCO5107. (blue: wild type; red: SC*glnK*-3 mutant)

SCO0259 show a very strong response to glutamate depletion in the wild type strain. Around this event both genes are upregulated by about 1.5 fold changes and then they are immediately downregulated to their initial level. In the wild type strain SCO5983 shows no reaction and SCO0259 is only slightly upregulated about 4 hours after the phosphate depletion event (see figure 4.24).

4.5.6. Ectoine biosynthesis gene cluster

An interesting phenomenon is observed for the genes of the ectoine biosynthesis cluster *ectABCD* (SCO1864-SCO1867). Around the time of glutamate depletion these genes are upregulated in the wild type strain and downregulated in the SC*glnK*-3 mutant (see figure 4.25). In the wild type strain about 3h after the upregulation the expression decreases. In the SC*glnK*-3 mutant the reaction of the *ect* genes is slower.

4.5.7. Genes involved in antibiotic biosynthesis

Although the production of actinorhodin (Act) is significantly lower in the SC*glnK*-3 mutant in comparison to the wild type strain, the expression levels of the respective biosynthesis genes (SCO5071-SCO5092) are quite similar at the end of the time series. However, in the wild type strain the upregulation of this gene cluster begins at the time of glutamate depletion about 35h after inoculation, whereas in the SC*glnK*-3 mutant it begins 2h after the depletion event and it is significantly steeper so that at about 42h after inoculation the expression level of the cluster is almost the same in both time series.

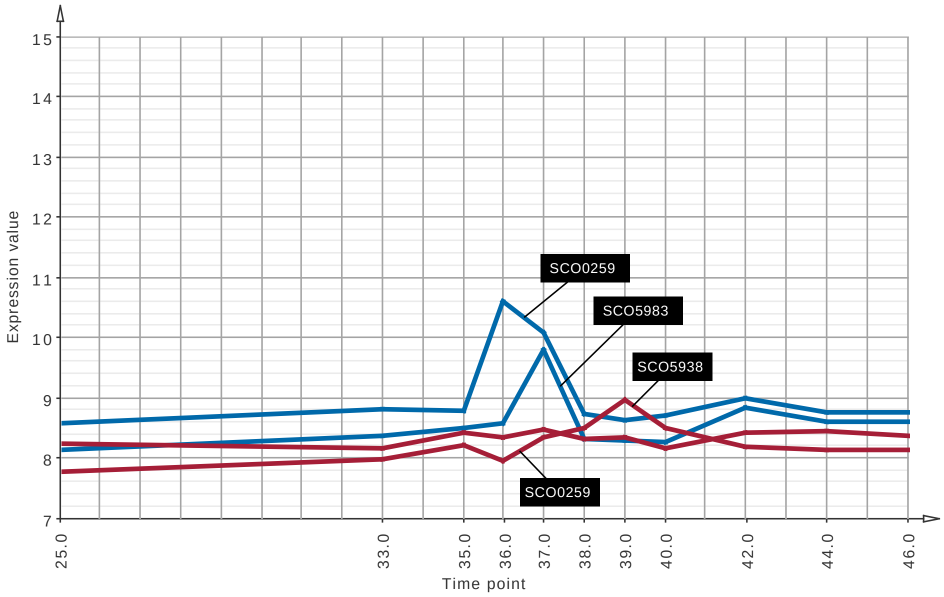


Figure 4.24.: Expression profiles of SCO5983 and SCO0259. (blue: wild type; red: SC*glnK*-3 mutant)

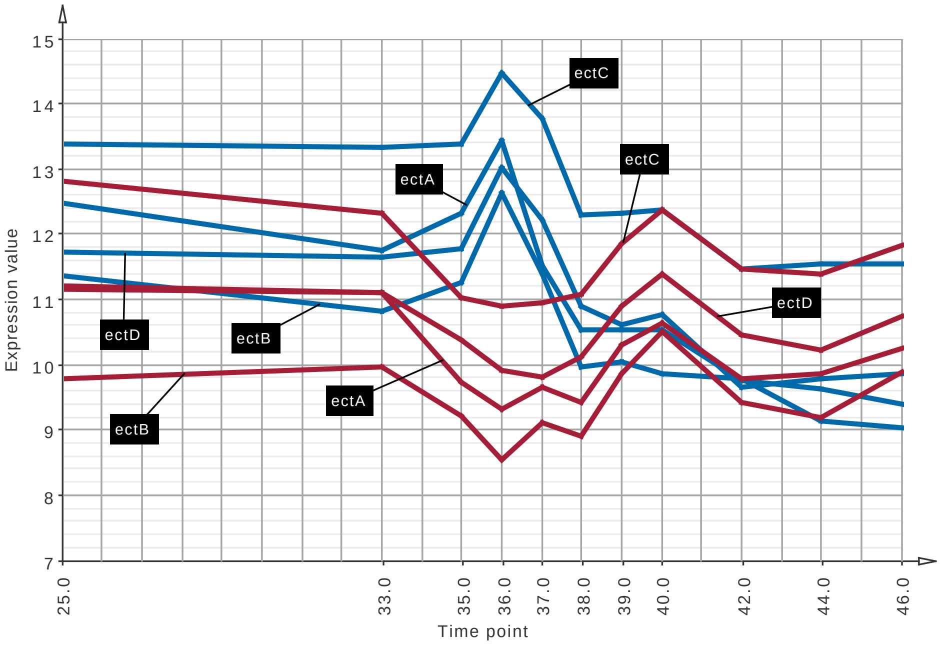


Figure 4.25.: Expression profiles of the *ectABCD* genes (blue: wild type; red: SC*glnK*-3 mutant)

In the SC*glnK*-3 mutant only a very small amount of undecylprodigiosin (Red) is produced. Interestingly, the expression profiles of the Red biosynthesis gene cluster do not differ considerably between the two time series.

4.5.8. Conclusions

The results presented here, represent the first high-resolution time series transcriptome study comparing a *Streptomyces coelicolor* wild type strain and a SC*glnK*-3 mutant strain under glutamate limited conditions. The comparative analysis of gene expression data of this level of detail bears several challenges as biologically relevant differences in expression patterns have to be distinguished from random fluctuations on a global scale. Here, these challenges have been successfully addressed by combining automated screening approaches with comparative visual analytics. By this, the effect of the used glutamate limited growth medium on the two strains could be studied in great detail.

Interestingly, most genes involved in nitrogen metabolism show similar expression profiles when comparing the wild type strain to the SC*glnK*-3 mutant or only slight changes can be observed. One possible explanation could be that the glutamate limitation of the medium was not as effective as expected. To further elucidate the role of GlnK as a regulator, more experiments using different growth media have to be conducted and analysed in a comparative manner as it has been presented here. Additional experiments will also help to further explore the interesting and in many cases unexpected expression patterns observed in this study, such as the dynamic expression behaviour of the ectoine biosynthesis genes.

5. The SuperGenome: A new representation of multiple whole-genome alignments

In the fields of genomics and transcriptomics comparative analyses between different species, strains or individuals are of major importance. Due to next generation sequencing technologies data can be produced in single nucleotide resolution and comparative methods have to operate at this level, for example in the context of comparative analyses of single nucleotide polymorphisms or transcription start sites.

In this chapter I will describe the SuperGenome, a concept that forms the fundamental basis of these comparative analyses. The SuperGenome is a novel way to represent multiple whole-genome alignments to provide a common coordinate system allowing for an efficient handling of comparative data. The flexibility of this approach is demonstrated by its application to the field of whole-genome alignment visualization.

5.1. The SuperGenome algorithm

There is a significant difference between multiple sequence alignments of short sequences (such as genes) and whole-genome alignments. In standard sequence alignments it is guaranteed that the order of the characters in the aligned sequences is equal to the order in the original sequences. From the perspective of sequence evolution this means that sequence information might be inserted or deleted or single characters might be changed, but the possibility that parts of the sequence might interchange their positions is not considered.

This assumption is true in most cases when dealing with short sequences and most alignment algorithms that guarantee an optimal solution would become computationally infeasible without this assumption, as dynamic programming would not be applicable any more.

However, in a whole-genome context the occurrence of rearrangements has to be considered [39, 148]. This not only includes translocations, where genomic regions change their location but also inversions, where genomic regions are replaced by their reverse complement. These events prevent the application of sequence alignment in the classical sense. However, it is still assumed that the regions that are affected by rearrangements are of a significant size and that therefore a standard sequence alignment is still possible within a local context.

The alignment of the complete genomes is then represented by a set of such locally collinear alignments, which are often referred to as *blocks*. In this context collinearity

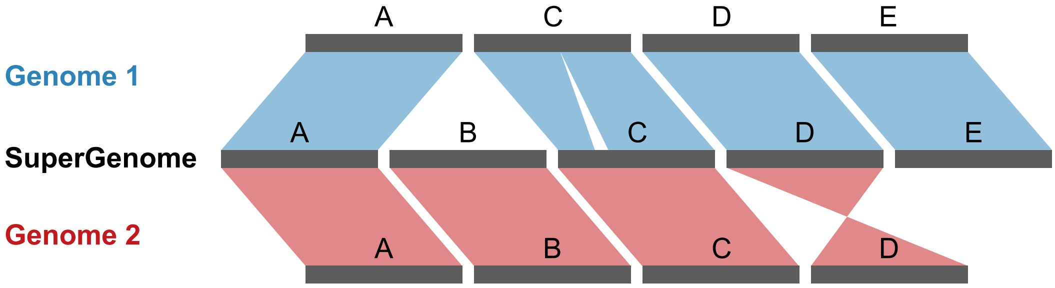


Figure 5.1.: Schematic representation of the SuperGenome concept. A SuperGenome is constructed for a set of aligned genomes (here *Genome 1* and *Genome 2*). The SuperGenome consists of all regions that are contained in at least one of the aligned regions. For example block *B* is only contained in *Genome 2* and block *E* is only contained in *Genome 1*. Thus, none of the aligned genomes contains all blocks and could serve as a proper reference for comparative analyses. The SuperGenome as a meta reference contains all regions and calculates mappings between the global SuperGenome coordinate system and the coordinate systems of the individual genomes (here denoted as colored ribbons). This high-resolution mapping allows for modelling small insertions or deletions (block *C*) and genomic rearrangements (block *D* represents an inversion between the two genomes).

refers to global gapped alignments by which the order of nucleotides in each of the aligned sequences is preserved. Programs computing whole-genome alignments of this form are for example `Mauve` [37, 38], `Mugsy` [9] and `TBA` [22]. Although solving the problem of multiple whole-genome alignments, this approach lacks some features of classical alignments such as an unambiguously defined alignment coordinate system. However, an unambiguous coordinate system is a necessity for comparative analyses.

Various programs for the comparative analysis and visualization of multiple genomes have been published (e.g., the `VISTA` program suite [48], `CGAT` [158] or `GECO` [86]; see also [5] for a review). Most solutions, however, are based on the selection of a reference genome especially for visualization. The necessity to define a specific reference can be problematic. For elements that are located in regions of other genomes that cannot be aligned to the chosen reference it is not possible to assign any alignment coordinates. Furthermore, the alignment coordinate system changes with the selected reference. As repeated analyses with altering reference genomes are infeasible in most cases, a concept of a meta reference is needed that represents all genomes that are compared in a study.

For this the SuperGenome concept has been developed in this dissertation in order to bridge the gap between whole-genome alignments that model rearrangements by producing a set of collinearly aligned regions and methods relying on a clearly defined coordinate system for comparative analyses, visualization or other applications. The SuperGenome is independent of a reference sequence and explicitly includes unaligned regions. A schematic representation of the SuperGenome concept is depicted in figure 5.1.

The SuperGenome was developed on `Mauve` alignments but the concept is applicable to all aligners producing a set of alignment blocks. The length of the SuperGenome is defined by the overall length of all blocks including unaligned regions. The core implementation of the SuperGenome consists of two integer arrays for each genome in the multiple alignment. One array stores for each genomic position the corresponding position in the SuperGenome, the second array stores for each position of the SuperGenome the respective position in the genome (the *inverse* mapping). Here, 1-based indexing applies, where an entry of the value 0 indicates that the respective SuperGenome position has no representative in the respective genome. Thus, this position is contained in one or more other genomes but could not be aligned to this genome. A negative value indicates a mapping to the respective position but on the other strand, thus, representing an inversion.

5.1.1. Preprocessing

The mapping is calculated by parsing and processing the alignment information in each alignment block. For this the blocks are first sorted according to their location in the aligned genomes. Though the functionality of the SuperGenome is completely independent from the order of the blocks, they have to be processed in some order and therefore it was decided to chose an ordering strategy that produces an order close to that in the aligned genomes. As the order of the blocks differs from genome to genome, the genomic orders are considered with different priority, which is determined by the order of the genomes in the alignment. This means that the first genome in the alignment has the highest priority. Thus, the order of all blocks that appear in the first genome is determined by their position in this genome. The order of all blocks that do not appear in the first but in the second genome of the alignment is determined by their position in the second genome, and so on. In this way the order of aligned regions in the SuperGenome is depending on their order in the aligned genomes with the first aligned genome having the most influence.

5.1.2. Coordinate mapping

In the next step the coordinate mapping is determined. For this each block is processed as follows. For each position i in each aligned sequence that is contained in the block the algorithm calculates at first the respective position in the source genome $g(i)$ by using the genomic start or end coordinate of the block ($g_{blockStart}$, $g_{blockEnd}$). If the sequence was taken from the forward strand the start coordinate ($g_{blockStart}$) is used as offset and for sequences taken from the reverse strand of the respective genome the end coordinate ($g_{blockEnd}$) is used.

The offset combined with the position in the sequence excluding gaps (i') defines the genomic coordinate for the forward strand ($g(i) = g_{blockStart} + i' - 1$) and the reverse strand ($g(i) = g_{blockEnd} - i' + 1$), respectively. The position in the block itself is simply given by the position in the sequence when gaps are included (i). The respective position in the SuperGenome $super(i)$ is then calculated by the position in the block added to the start coordinate of the block in the

SuperGenome ($super_{blockStart}$), which is the sum of the lengths of all previous blocks plus 1 ($super(i) = super_{blockStart} + i$).

The respective genomic coordinate (including strand modifier) is then stored at this index in the SuperGenome mapping array for that genome and the SuperGenome position is stored in the genomic array, which points to the SuperGenome, respectively.

These two mapping arrays, which are calculated for each genome form the core data structure of the SuperGenome. This structure is utilized for all operations that directly act on the SuperGenome.

5.1.3. The SuperGenome Interface

Any direct operations on the SuperGenome data structure require detailed knowledge of how this structure represents the multiple whole-genome alignment and how this information has to be processed. Thus, the SuperGenome implementation contains a set of wrapper functions performing various SuperGenome-based coordinate and sequence transformations that allow the programmer to make use of the SuperGenome's functionality without full knowledge of the underlying data structure.

Position mapping The most basic transformations concern the mapping of single positions. The function `getPosInGenome(String genomeID, int superGenomePos)` takes a genome ID and a SuperGenome coordinate as input and returns the respective genomic coordinate. If there is no mapping of this specific SuperGenome position into the respective genome, a value of 0 is returned. If the respective genomic coordinate is located inside a region that is inverted in relation to the SuperGenome a negative value is returned.

Respectively, the function `getPosInSuperGenome(String genomeID, int genomePos)` takes a genomic position as input and returns the corresponding SuperGenome position using the same encoding for inversions. A value of 0, however, will never be returned by this function as there are no genomic positions which are not represented by the SuperGenome.

When a SuperGenome position cannot be mapped to a genome, it might be useful to get the closest position, where a mapping is possible. For this the function `getNextMappingPosInGenome(String genomeID, int superGenomePos)` can be used. It takes a genome ID and a SuperGenome position as input and searches for the SuperGenome position that is closest to the input position and can be mapped to the respective genome. Then the corresponding genomic position is returned.

Mapping position specific data RNA-seq expression data is often provided in the `wiggle` format, which consists of two columns, where the first contains the genomic coordinate and the second contains the expression value for that position. There are only entries for positions, where expression has been detected.

These wiggle tracks can be the basis for measuring gene expression or performing other types of analysis like detecting transcription start sites (TSS) as described in chapter 6. The function `superGenomifyXYtrack(String genomeID, double xyTrack[])` takes such a track as input, which relates to the coordinate system of a specific genome, and transfers it into the coordinate system of the SuperGenome. This for example allows for the comparative visualization of RNA-seq data from different organisms as they can then be visualized within the same coordinate system, for which a standard genome browser can be used, if the SuperGenome is used as the reference.

Mapping genomic annotations In addition to RNA-seq data genome annotations can be mapped into the SuperGenome coordinate system or from the SuperGenome into individual genomes. This includes intervals like genes (`superGenomifyGenes, genomifySuperGenes`) and single-nucleotide annotations like TSS (`superGenomifyTSS, genomifySuperTSS`).

When genes from different genomes are mapped to the SuperGenome they can be directly compared using their SuperGenome coordinates. E.g., if two genes from different genomes overlap in the SuperGenome, this means that these genes or parts of the genes are aligned in the alignment the SuperGenome is based on. Therefore, by the simple calculation of interval intersections assumptions about the similarity and thus homology of genes can be made, which can be used as the basis for a SuperGenome-based ortholog detection. In section 6.1.8 this mechanism is used to compare expression values of orthologous genes.

The TSS detection algorithm described in chapter 6 makes also use of this functionality to associate TSS to each other that have been detected in different genomes.

Sequence transformation When investigating potential orthologs or promoter regions of associated TSS a direct sequence comparison is necessary. To accomplish this the function `superGenomifyFASTA` can be used to map a genomic sequence into the SuperGenome coordinate system. To SuperGenome positions to which there is no mapping from the respective genome the gap character (-) is assigned. These sequences can also be used for visualization in an alignment viewer or in a genome browser alongside RNA-seq data, for example.

In addition, a SuperGenome consensus sequence can be generated by using the function `superGenomeConsensus`. This is done by determining for each SuperGenome position the most abundant nucleotide that is found at the respective position in all genomes to which this SuperGenome position can be mapped. The SuperGenome consensus does not contain any gaps.

Alignment statistics The SuperGenome provides several functions to calculate basic alignment statistics. These encompass the number of perfectly matching alignment columns (`getPerfectColCount`) or the number of deletions and insertions for each genome (`getDelCountMap, getInsCountMap`). Here, deletions are defined as columns, where one genome contains a gap and at least two other genomes are aligned at that position. An insertion is defined as an alignment column that only contains one character and otherwise only gaps.

Furthermore, the length of the SuperGenome, which corresponds to the length of the complete alignment, can be retrieved by using the function `getAlignmentLength`.

Exporting the SuperGenome mapping The core of the SuperGenome, which is the coordinate mapping between the individual genomes and the SuperGenome coordinate system can be exported to a plain text file for external processing. The exported file consists of one row for each SuperGenome position and each row consists of the SuperGenome coordinate and the respective coordinates in the individual genomes. If a SuperGenome position cannot be mapped to a genome the entry for this genome is '0'. Inversions are represented by negative coordinates.

5.2. GenomeRing: alignment visualization based on SuperGenome coordinates

The SuperGenome concept was the basis of a new genome alignment visualization approach [64].

The field of genome visualization is increasingly dynamic as more and more genomic data become available. There are various genome visualization approaches (see Nielsen *et al.* for a review [108]), which in some cases aim at a comparative visualization of multiple genomes.

However, many challenges involved in this field still remain to be overcome. Many existing visualization techniques, for example, do not address the problem of deviating coordinate systems of comparatively visualized genomes. Regions of similarity are often highlighted by ribbons or indicated by color coding. This, however, quickly leads to visual clutter and, in addition regions of similarity are often not visually aligned. The absence of a common coordinate system also makes the comparative visualization of genome annotations very challenging. Here, GenomeRing is presented, which specifically focuses on the comprehensive visualization of differences and similarities between genomes on the basis of a SuperGenome coordinate system based on a multiple whole-genome alignment. With this combination of the SuperGenome concept and a circular multiple genome visualizer GenomeRing won the *Most Creative Algorithm Award* of the *Illumina iDEA Challenge 2011.*

5.2.1. Preprocessing

In a first step the SuperGenome is used to generate a set of blocks, which represent genomic regions that are either shared by two or more genomes or that are unique to single genomes. To accomplish this, the block information as provided by the aligner is further refined. Each sequence in each alignment block is scanned for gap regions that are longer than a user defined threshold (*minBlockSize*). These regions represent loci that are missing from one or more genomes. At the start and the end coordinates of these regions break points are annotated, which potentially give rise to borders of new blocks. The break point sets of all sequences in an alignment block are then merged and subblocks are generated for each consecutive pair of break points which result in subblocks not shorter than the minimal block size (*minBlockSize*). In a last step neighboring subblocks with the same genome composition are merged (i.e., subblocks representing regions in the same set of genomes). The final set of all subblocks is then subject to visualization with GenomeRing.

5.2.2. Layout of GenomeRing

GenomeRing visualizes the set of SuperGenome blocks, which represent the multiple whole-genome alignment, with a circular layout. It consists of two concentric circles, one for the forward and one for the reverse direction, which allows for the indication of inversions. An overview of GenomeRing's visual components is presented in figure 5.2.

The SuperGenome blocks are laid out on the two circles with each block being visually represented on both of them. For each genome there is a colored path that traverses blocks that are contained in this genome and that skips blocks that are not contained. By this the path connects the blocks according to their order in the respective genome. If the path traverses a block on the inner ring the respective region is inverted in comparison to the genomes whose paths traverse the block on the outer ring. Start and end of each genome are indicated by small colored flags.

In GenomeRing, for each block it can be quickly identified in which genomes it appears by visually evaluating its color composition, i.e., the colors of the paths that traverse this block.

This kind of layout allows for an easy identification of genomic regions where the aligned genomes are similar or where they vary or which are unique for a certain genome. The structure of each individual genome is preserved in the visualization by the respective path. Therefore, the actual order of the SuperGenome blocks on the circles is arbitrary and can be optimized to enhance visual clarity. For this several heuristics have been integrated that reorder the SuperGenome blocks to minimize the number and the length of the jump edges, i.e., the number of events in which a block has to be skipped by a genome's path.

5.2.3. Uncovering differences between genomic architectures

One of GenomeRing's main applications is the discovery of large-scale deletions or insertions. These can be due to genomic islands, which are caused by horizontal transfer, or prophages or plasmids that have been integrated into the chromosome. Such regions are of major interest because they potentially contain virulence factors in the case of pathogenicity islands or genes involved in drug resistance in the case of antibiotic resistance islands.

To demonstrate how GenomeRing can be used to identify these kind of regions, it was applied to four *Campylobacter jejuni* strains (RM1221, NCTC11168, 81-176, 81116). This is a Gram-negative food-born pathogen that is one of the major causes of gastroenteritis [145].

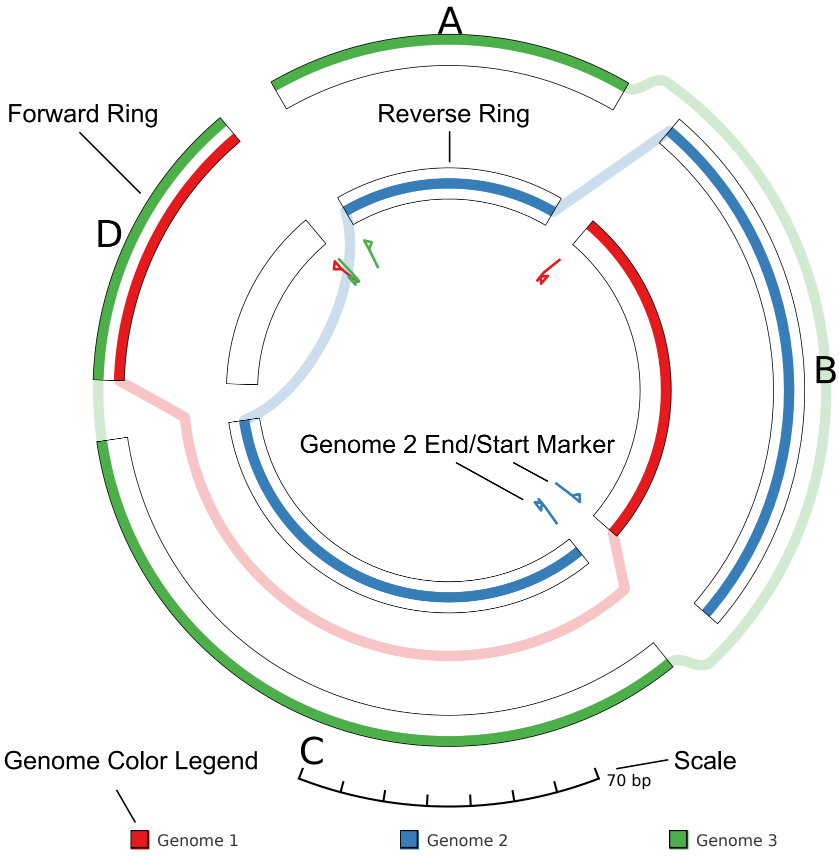


Figure 5.2.: GenomeRing visualization of an artificial example involving three genomes. The SuperGenome consists of four blocks *A*, *B*, *C* and *D*. The SuperGenome blocks are laid out on two concentric circles (rings). For each genome a colored path connects the blocks in the order as they appear in that genome. Blocks that are not contained in the genome are skipped. The path traverses blocks that are contained in the genome either on the outer ring or the inner ring. By this, inversions between the genomes in the outer ring and the genomes in the inner ring are visualized. The start and end of each genome are depicted as small colored flags. In this example *Genome 1* consists of blocks *D* and *B*, *Genome 2* consists of blocks *B*, *A*, *C* and *Genome 3* consists of blocks *A*, *C*, *D*. For each block it can be easily seen in which genomes it occurs by evaluating its color composition. For example, block *D* is contained in genomes *1* and *3* (red and green) while block *A* is contained in genomes *2* and *3* (blue and green). In block *A Genome 2* (blue) is shown on the inner ring and *Genome 3* (green) is shown on the outer ring. Thus, block *A* represents an inversion between genomes *2* and *3*.

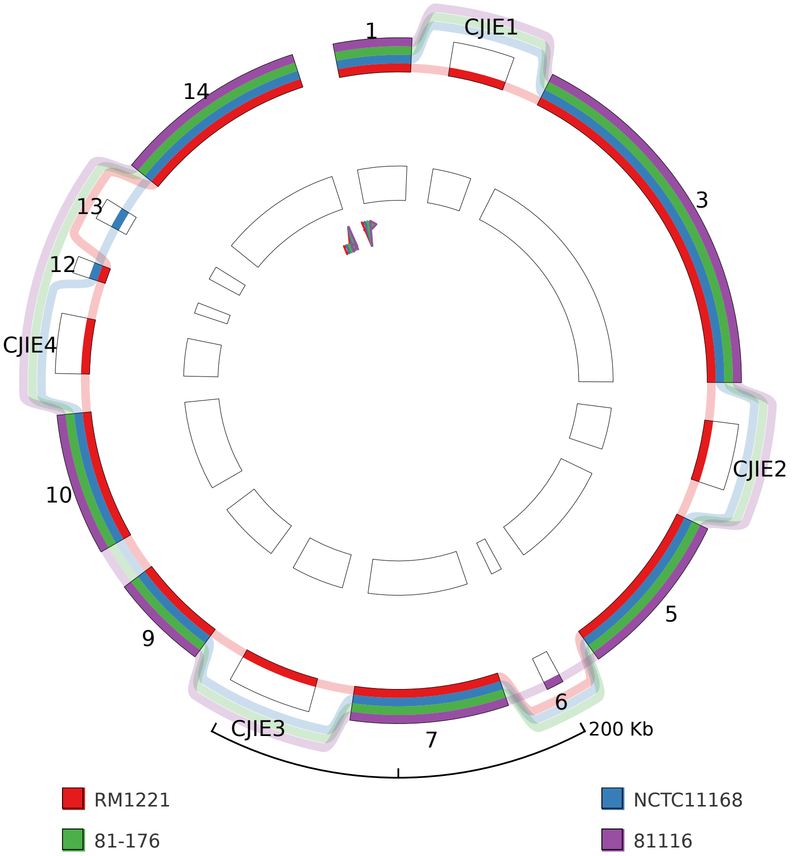


Figure 5.3.: GenomeRing visualization of an alignment of the four *Campylobacter jejuni* strains RM1221, NCTC11168, 81-176 and 81116. Four of the blocks indicate insertions in *C. jejuni* RM1221 and represent genomic islands (CJIE1-4). The parameter for the minimal block length was set to 10 kb. The inner ring is empty in this view as there are no inversions between the genomes. [64]

The SuperGenome was generated with a minimal block size of 10 kb. After subblock generation the resulting SuperGenome consisted of 14 blocks, which were subject to the GenomeRing visualization (see figure 5.3). For each block the color pattern indicates in which genomes the respective block is contained. Thus, a block which is traversed by all genome paths because it is conserved in all genomes shows the full color pattern. In this example most blocks can be found in all genomes. However, there are four large blocks representing insertions in *C. jejuni* RM1221. These blocks can be easily identified in the visualization as they are traversed by only one genome path (that of RM1221) while it is skipped by all other paths. The respective regions correspond to, so-called, *Campylobacter jejuni*-integrated elements (CJIEs) [47, 116]. CJIE1 is a Mu-like phage and therefore also called CMLP1 (*Campylobacter* Mu-like phage 1). CJIE2 and CJIE4 also contain phage-related proteins, whereas CJIE3 is a putative integrated plasmid.

5.2.4. Integrating genome annotations

One major purpose of the SuperGenome is to allow for a consistent assignment of coordinates to genome annotations even in the presence of insertions, deletions and genomic rearrangements. This functionality is also used here to display genome annotations in GenomeRing. GenomeRing is implemented in the transcriptome analysis software Mayday [14], which, in addition, allows for an integration of gene expression data, for example.

This is demonstrated by the application of the SuperGenome and GenomeRing to an alignment of the three *Helicobacter pylori* strains 26695, J99 and P12. In addition, gene expression data for *H. pylori* 26695 was analysed with Mayday and the results were visualized in GenomeRing. *Helicobacter pylori* is a Gram-negative pathogen populating the human stomach. It can cause gastritis and also gastric cancer [34]. Large parts of the human population are infected with this bacterium, which is, however, asymptomatic for most individuals. Understanding the mechanisms that are responsible for its pathogenicity is therefore of major importance in order to develop effective treatments. In 2010 Sharma *et al.* completed a very comprehensive transcriptomic study of *Helicobacter pylori* strain 26695 cultivating the organism under five conditions [142]. It was grown to mid-logarithmic phase (ML), under acid stress (AS), in contact with responsive gastric epithelial cells (AG) and non-responsive liver cells (HU), and in pure cell culture medium (PL).

The SuperGenome generation was applied to the alignment of the three strains with a minimal block size of 50 kb, thus, only showing large differences between the strains. This results in a SuperGenome that contains eight blocks. Two of these blocks are due to inversions between the strains 26695 and J99/P12 (see figure 5.4).

To integrate the expression data of the study by Sharma *et al.* it was loaded into Mayday and a z-score normalization followed by a k-means clustering of the expression profiles was performed. By this groups of genes were identified that are differentially expressed under a certain cultivation condition. For the visualization two large expression profile clusters were selected, where one was upregulated under acid stress (AS) and the other one showed an upregulation when the bacteria grew in contact with liver cells (HU). Different colors were assigned to the two clusters and the loci of the genes that are contained in the clusters were mapped into the path of *H. pylori* strain 26695 in the GenomeRing visualization using the color of the respective cluster. This provides a comprehensive overview of the location of genes that show differential regulation under a certain experimental condition.

GenomeRing easily shows that in many cases genes that are upregulated under the same cultivation condition, either AS or HU, are located in close vicinity. One such locus is highlighted in figure 5.4. These appear as stretches of loci that are visualized in the same color. The investigation of genes that are co-localized and co-expressed under specific conditions is of major interest as these genes are often involved in the same pathways and are therefore functionally related.

In order to enable the user to investigate identified regions of interest in more detail, GenomeRing can be linked to the genome browser that is integrated into Mayday.

Linkage to Mayday's genome browser The advantage of GenomeRing's integration in Mayday is that locus-specific expression data can be visualized and GenomeRing can be linked to Mayday's linear genome browser [150]. This allows for a selection of a specific region region in the GenomeRing visualization, which will then be displayed in the linear browser and can thus be investigated more in detail. An example is given in figure 5.5. In the linear browser the gene loci are visualized alongside a, so-called, heatmap track that indicates the gene expression values for the different cultivation conditions. Furthermore, the expression is also visualized in single-nucleotide resolution as wiggle tracks for the two conditions HU and AS.

There are two clusters of co-expressed genes which are located in the respective region. The larger cluster consists of genes that are upregulated when the organism grows in contact with liver cells (HU condition). The second cluster, which is located downstream, is smaller and contains genes that are upregulated under acid stress (AS condition). Most of the genes in the larger cluster are ribosomal proteins. Of the four genes that are contained in the smaller cluster two are annotated as hypothetical proteins and two as cation efflux system proteins (*czcA*), which are known to be activated at low pH levels [21]. The co-expression of the four genes suggests that the two proteins of unknown function are potentially involved in the same system or a similar protective mechanism.

At this point Mayday's genome browser can be further used for a more detailed analysis of the locus. The heatmap track in figure 5.5B, for example, shows that the genes in the two clusters are very specifically upregulated under condition HU and AS, respectively, while being at a quite low expression level under the other conditions. In addition, the single-nucleotide resolution RNA-seq data visualized as wiggle tracks can be employed for the characterization of the architectures of chromosomal genes clusters, e.g., in the context of TSS and operon prediction.

Thereby this applications demonstrates how the visualization concept of GenomeRing, which is based on the SuperGenome, can be used to gain an instant overview of similarities and differences of the investigated genomes, while its integration in Mayday allows for the inclusion of transcriptomic data analysis and visual-

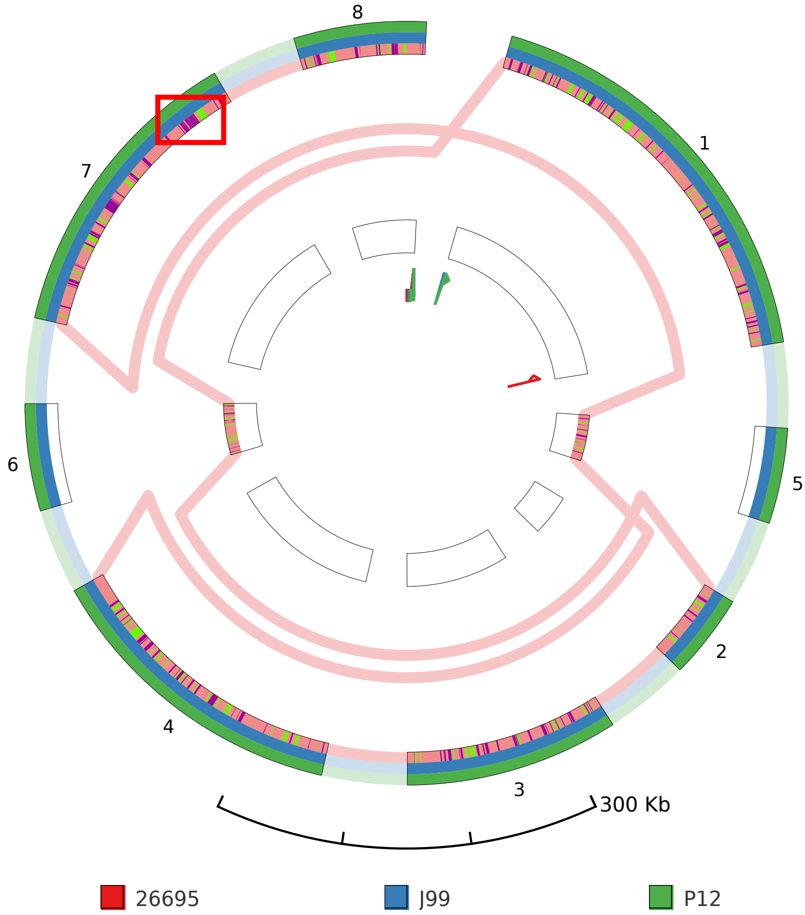


Figure 5.4.: GenomeRing visualization of an alignment of the three *Helicobacter pylori* strains 26695, J99 and P12. The parameter for the minimal block length was set to 50 kb. Blocks 5 and 6 represent two large inversions between strains 26695 and J99/P12. For strain 26695 gene expression data has been mapped into the respective path (red). Genes that are upregulated in condition HU are shown in purple while genes upregulated in condition AS are shown in green. The locus indicated by a red rectangle is shown in more detail in figure 5.5. [64]

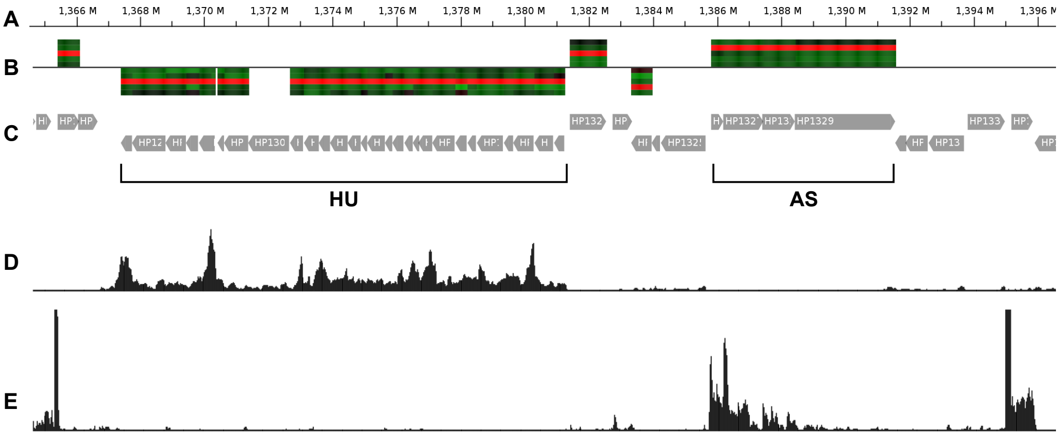


Figure 5.5.: Visualization of the locus highlighted in figure 5.4 in the genome of *Helicobacter pylori* 26695 using the linear genome browser integrated in Mayday. Five tracks are displayed: A, genomic coordinates for the genome of *H. pylori* 26695. B, expression value heatmap track for genes that are upregulated under conditions HU or AS (forward strand: above the baseline, reverse strand: below the baseline). The expression of all five conditions is shown in the heatmap (from top to bottom: AG, AS, HU, ML, PL). C, annotations of protein-coding genes. Co-localized genes that are upregulated under the same condition are highlighted by horizontal braces. D, RNA-seq wiggle track for the condition HU (reverse strand). E, RNA-seq wiggle track for the condition AS (forward strand). [64]

ization. Furthermore, the linkage with Mayday's linear genome browser enables the detailed inspection of relevant loci, which have been identified in the GenomeRing visualization.

6. Comparative prediction of TSS using the SuperGenome

RNA-seq data allows for unprecedented insights into the transcriptomic structure of an organism as expression is measured in single-nucleotide resolution. Thus, the importance of the automated analysis of RNA-seq data becomes evident as increasingly more and more data is generated not just from single transcriptomes but either from different conditions or from different organisms.

However, this leads to new challenges for the researcher and for computational methods as comparative analyses also have to be performed on single-nucleotide level.

One of the genomic features that can be deduced from RNA-seq data, for which comparable high-throughput methods did not exist before, are transcription start sites (TSS). Information on where exactly in the genome a transcript starts is of major importance in order to identify features involved in the regulation of transcription such as sigma factor binding sites or other transcription factor binding sites. In addition, accurate genome-wide TSS maps can assist in the identification and characterization of promoter regions or *cis*-regulatory elements that are part of the 5' UTR of the transcript, such as riboswitches.

However, an organism-specific global TSS map cannot easily be used for further comparative analyses, because even if TSS maps for other organisms exist, they are not directly comparable, as they refer to completely independent coordinate systems. TSS in different maps could be associated to each other for example by mapping of orthologous genes [81]. However, this strategy might lack accuracy as TSS in different organisms can only be roughly associated to each other by their location with respect to orthologs but not in single-nucleotide resolution. Also, this approach does not allow for a comparative analysis of TSS that cannot be related to known genes.

For these reasons it is desirable to directly perform a comparative detection and characterization of TSS on multiple organisms. In order to achieve a high accuracy and sensitivity TSS need to be detected and assigned to each other in single-nucleotide resolution. For this, the comparative analysis has to take place in the context of a global coordinate system based on the genomes included in the study.

To accomplish this an algorithm was developed that allows for a fully automated genome-wide annotation of TSS in a comparative manner by the integration of the SuperGenome (chapter 5). The approach is complemented by methods for the normalization of RNA-seq expression data and the automated classification of detected TSS. In order to provide these functionalities in a user friendly environment

TSSpredator was developed, which combines the methods in a common frame work with a graphical user interface.

6.1. The TSS prediction pipeline

The complete data processing pipeline resulting in the prediction of global comparative TSS maps consists of several steps. In the first step differential RNA-seq (dRNA-seq) data is read and normalized. The dRNA-seq technique was developed by Sharma *et al.* in 2010 [142]. Using this protocol two libraries are produced for RNA sequencing. One standard library remains untreated. A second library is treated with a terminator exonuclease that specifically degrades RNA fragments with a 5' monophosphate. By this, the 5' ends of primary transcripts, which carry a 5' triphosphate are enriched in this library. In order to distinguish real TSS from RNA processing sites this treated library is compared to the untreated library and only sites that appear to be enriched in the treated library can be considered to be TSS. The dRNA-seq technique is illustrated in figure 6.1.

After normalization of the data TSS candidates are detected in the replicates of all data sets and replicates are compared to eliminate candidates that could not be reproduced. In the next step the TSS candidates of the different data sets are associated to each other in order to decide for each TSS in which of the data sets it could be detected. For the comparative analysis among different strains or species the SuperGenome approach is used in this step. Finally, the detected TSS are classified with respect to their location relatively to annotated genes and the results are presented in the form of a comprehensive MasterTable consisting of all predicted TSS and detailed information on each element. An overview of the TSS prediction pipeline is presented in figure 6.2. In the following all steps of the pipeline are described in detail.

6.1.1. Normalization

The first data processing step following an RNA-seq experiment is read mapping. The TSS detection method described here is not working on the mapping data directly but on coverage graphs in single nucleotide resolution that are derived from these data. These graphs (also called *wiggle* graphs) basically consist of a value for each genomic position indicating the number of mapped reads covering the respective position. As the mapping is strand-specific this results in two graphs per library, one for the forward and one for the reverse strand.

The graphs are usually normalized by the complete number of reads that could be mapped from this library. However, this number is often biases by only a few strongly expressed transcripts [43, 130]. These can be ribosomal RNAs, as the efficiency of the rRNA depletion protocol might vary between libraries. For this reason an additional normalization of the dRNA-seq graphs is conducted prior to TSS detection. This is done by performing a percentile normalization, which is more robust against the variation of very strongly expressed genes than using the total number of mapped reads as a normalization factor. For this the 90th percentile of all expression values

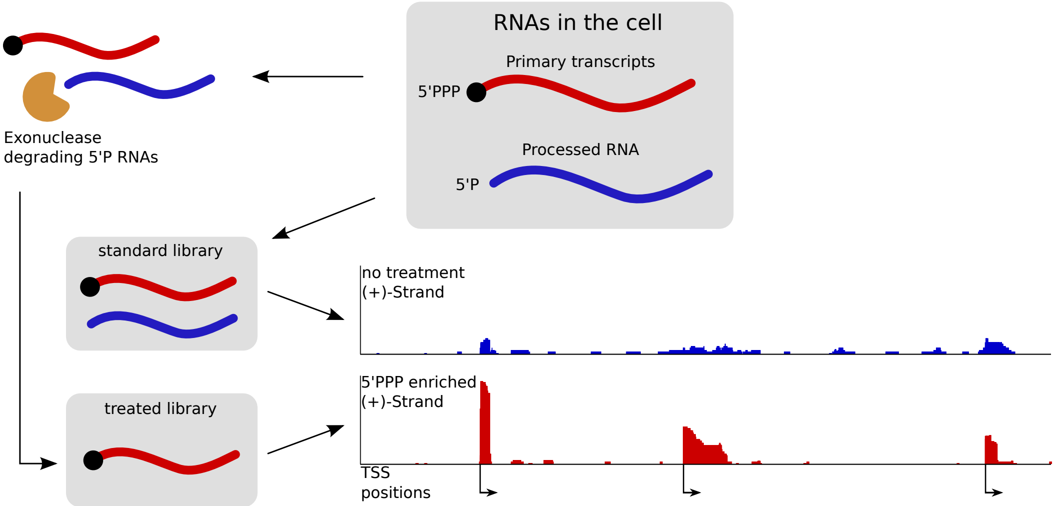


Figure 6.1.: Illustration of the differential RNA-seq (dRNA-seq) protocol. Two RNA sequencing libraries are produced, one that is untreated and one that is treated with a terminator exonuclease that degrades RNAs with a 5' monophosphate in order to enrich the 5' ends of primary transcripts that carry a 5' triphosphate. Both libraries are sequenced and the sequencing data are compared to distinguish TSS from RNA processing sites.

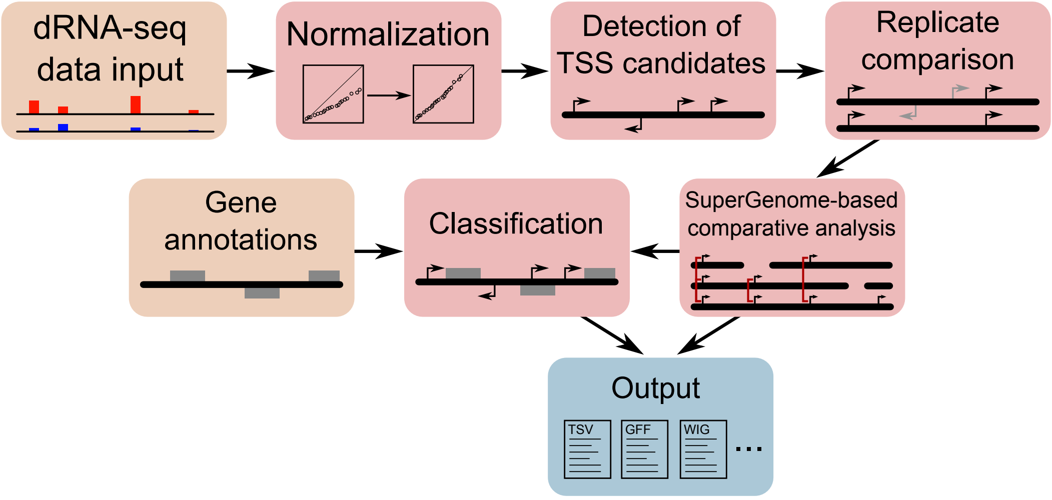


Figure 6.2.: Basic steps of the TSS prediction pipeline. The dRNA-seq input data is read and normalized. TSS candidates are detected in the replicates of the different data sets and the results are compared in order to eliminate irreproducible sites. The SuperGenome approach is employed to associate TSS candidates from different genomes to each other. After classification with respect to annotated genes all results such as the TSS MasterTable and supplemental data are generated.

is calculated and used as the normalization factor. The factor is calculated from the treated library, but it is applied to both, the treated and the untreated library. Thus, the enrichment factors are not changed during this normalization step. After the dRNA-seq graphs of libraries have been normalized, all expression values are multiplied by the minimal normalization factor in order to restore the original data range.

This normalization procedure actually makes no assumptions about the normalization state of the input data as the result is independent of any factor that have have been applied as a normalization factor earlier. However, still a linear normalization is used, which might not be sufficient if non-linear effects occur. A comparison of TSS expression height distributions after normalization between 4 RNA-seq libraries of 4 different *Campylobacter jejuni* strains (see chapter 7 for details) is shown as a Q-Q plot matrix in figure 6.3. For several libraries non-linear effects are evident. However, reasonable expression height thresholds for the annotation of TSS are between 5 and 10 reads. In this interval the normalization strategy presented here seems to be sufficient as pronounced non-linear effects are only observed for much higher expression levels.

Another important property for TSS prediction is the enrichment factor, i.e., the factor by which the expression value in the treated library is higher than in the untreated library. It has to be considered that the efficiency of the enrichment procedure directly influences the number of detectable TSS. Variations between the enrichment rates of different libraries biases the comparative analysis. In figure 6.4 the distributions of enrichment factors of predicted TSS are compared between 4 different *C. jejuni* strains. Here, only the normalization method described above was applied. The enrichment rates differ significantly between the strains. E.g., in strain NC_009312 the enrichment strength was about twice as high compared to strain NC_009839.

To account for this effect an additional normalization method was integrated. For this a preliminary prediction of TSS is performed for each pair of treated and untreated library, which uses fixed thresholds of 0.1 for the minimal step height and 1.5 for the minimal step factor (see 6.1.2). Other properties are not evaluated. The resulting TSS set is then used to determine the median enrichment factor for the respective library pair. Taking the library pair with the strongest enrichment as reference these values are used to determine for each pair the normalization factor that is necessary to achieve the same rate as the reference. This factor is then applied to the dRNA-seq graphs of the respective untreated library. In figure 6.5 the same comparison as described above is shown but with this additional normalization applied to the data. As for the expression heights there seem to be additional non-linear differences between the libraries. However, a reasonable threshold for the minimal enrichment factor will presumably be smaller than 10 in any case and the normalization compensates for all significant effects in an interval between 0 and 20.

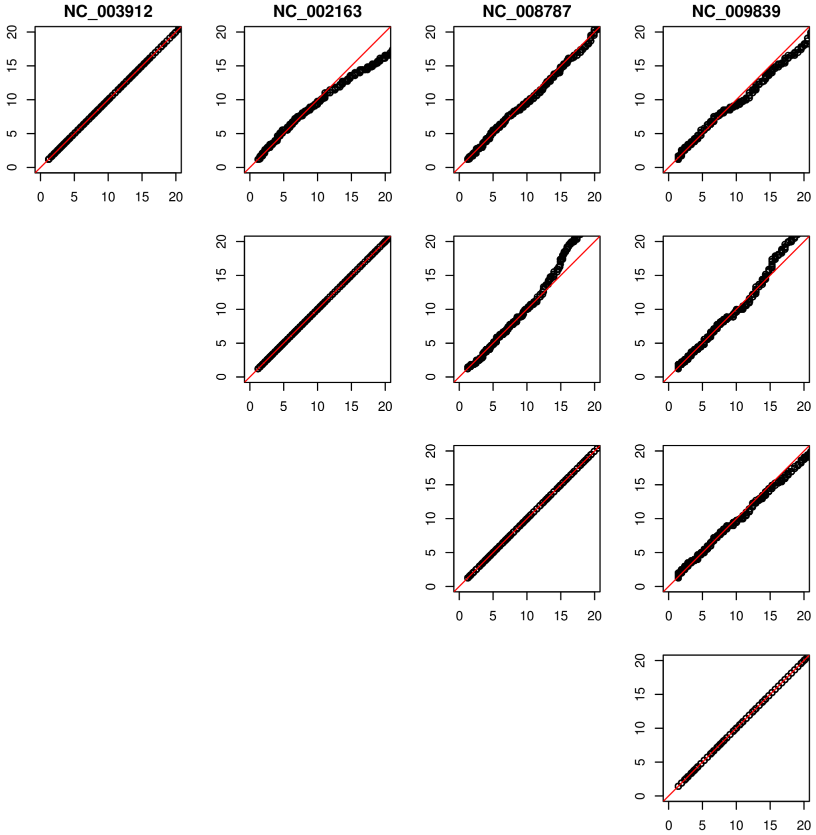


Figure 6.3.: Q-Q plot matrix comparing the distributions of TSS expression heights after normalization between 4 *Campylobacter jejuni* strains. Only the interval between 0 and 20 reads, which is relevant for the threshold of the TSS prediction method is shown.

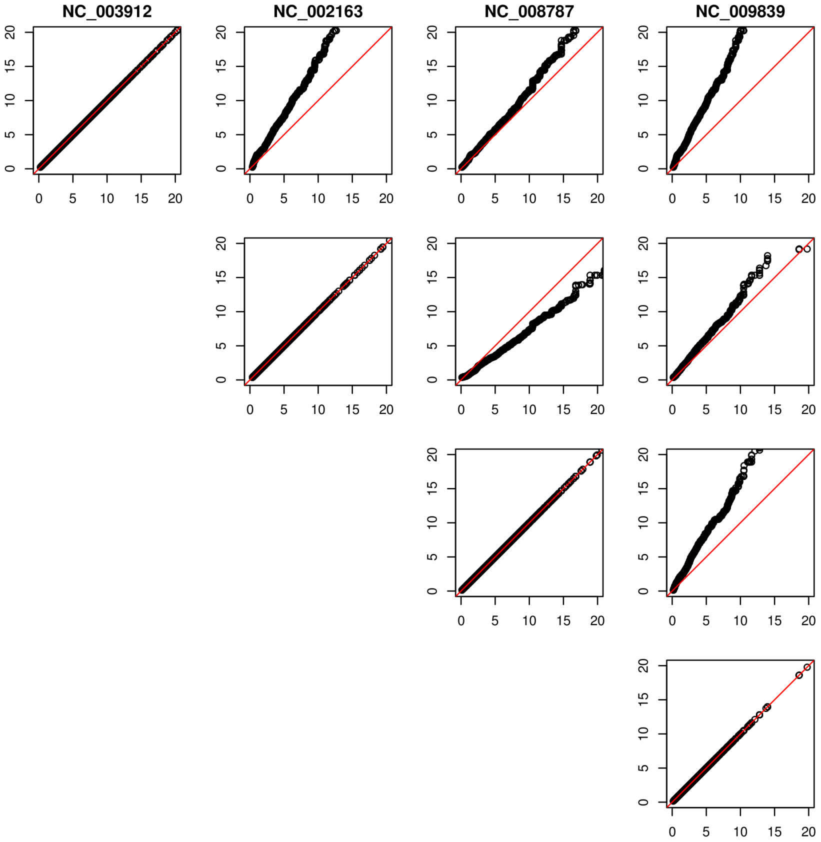


Figure 6.4.: Q-Q plot matrix comparing the distributions of TSS enrichment factors without additional normalization between 4 *Campylobacter jejuni* strains. Only the interval between 0 and 20 reads, which is relevant for the threshold of the TSS prediction method is shown.

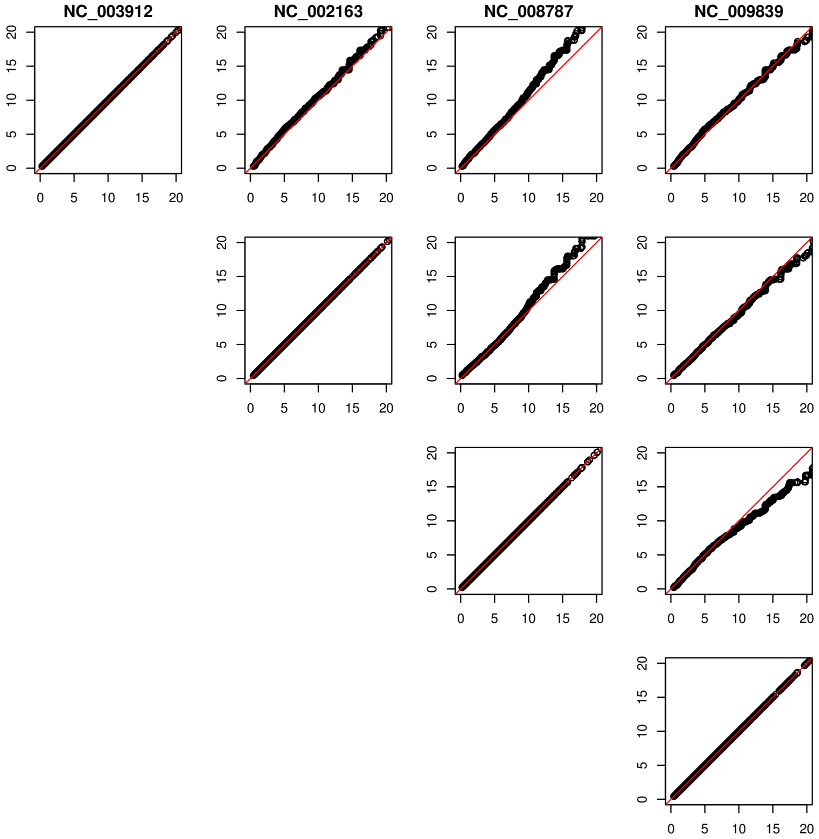


Figure 6.5.: Q-Q plot matrix comparing the distributions of TSS enrichment factors with additional normalization between 4 *Campylobacter jejuni* strains. Only the interval between 0 and 20 reads, which is relevant for the threshold of the TSS prediction method is shown.

6.1.2. Basic TSS detection procedure

The basic prediction of TSS is taking place on three different levels. The initial prediction is performed on the replicate level for each replicate separately. On the genome level the results from all replicates of the respective genome are combined to get an individual prediction for that genome. Finally, on the SuperGenome level the TSS of the individual genomes are assigned to each other as described in the next section (6.1.3).

For the initial prediction of TSS in each replicate the dRNA-seq graphs are processed as follows: In a first step the algorithm localizes positions in the graph of the treated library where a significant number of reads start, because these loci are potential transcript starts. More precisely, for each position i in the graph the absolute change of the expression height in comparison to the previous position $e(i) - e(i-1)$ is calculated, where $e(i)$ is the expression height at position i (Figure 6.6). However, this criterion alone would not be able to consider the local background of the TSS candidate. Therefore, the factor of height change $e(i)/e(i-1)$ is calculated. Using default thresholds the factor has to be greater than or equal to 1.5 and the absolute height has to be greater than or equal to 0.1. Note that the latter is a relative value with respect to the normalization factor, which is the $90th$ percentile by default. Thus, the default threshold for the expression height is $0.1 \cdot 90th$ percentile. By relating the threshold to the normalization factor the necessity to adapt it to the data range is avoided. In addition to those two thresholds it is evaluated for how many base pairs the expression height stays above the threshold. This value also has to exceed a certain threshold, which depends on the length of the reads that have been produced during the RNA-seq experiment. This is to prevent only partially mapping reads from being detected as a TSS. Additionally the enrichment factor at the position of the candidate is calculated as the ratio of the expression values of the treated and the untreated library ($e_{treated}(i)/e_{untreated}(i)$). However, the factor is not evaluated in this step. The default values for the thresholds have been determined by evaluating the method on a global set of manually annotated TSS from a dRNA-seq study by Sharma *et al.* in 2010 [142].

This process is performed for both strands separately. In the next step the TSS candidate sets of all replicates of an experiment are compared to evaluate in how many replicates a TSS candidate was detected. For this comparison TSS candidates from different replicates are assigned to each other if their position does not differ by more than 1 bp. TSS are then summarized across replicates so that there is one final set of TSS for the respective genome. During this step the position of the genomic TSS is determined by the respective replicate TSS with the highest expression. The absolute expression height, factor of height change and enrichment factor of the genomic TSS are determined by taking the respective maxima over all replicate TSS that contributed to the genomic TSS. By default a TSS candidate has to be detected in at least one replicate to be kept. However, in order to increase specificity this value can be increased.

At many loci where a transcript starts it can be observed that the expression height of the dRNA-seq graph increases in several steps, which might lead to an annotation of several TSS candidates. Although there are many cases where alternative TSS can

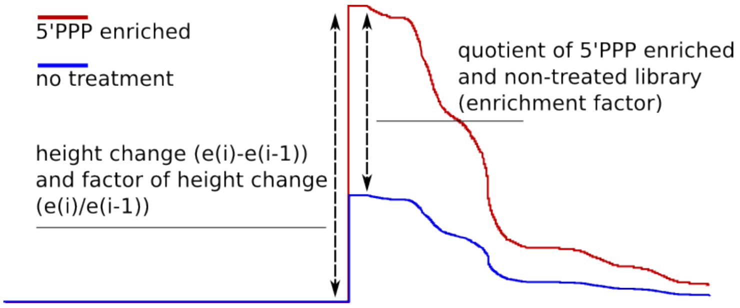

5'PPP enriched
no treatment
height change (e(i)-e(i-1)) and factor of height change (e(i)/e(i-1))
quotient of 5'PPP enriched and non-treated library (enrichment factor)

Figure 6.6.: Schematic representation of basic TSS detection criteria. The expression graphs from the treated (red) and untreated (blue) library are the basis for the TSS detection procedure. The most important parameters that are considered during the process are the absolute expression height at the position of the potential TSS $(e(i) - e(i-1))$ and the factor of height change at that position $(e(i)/e(i-1))$, where $e(i)$ is the expression value at position i. Additionally, the enrichment factor at the same position is taken into account (i.e. the ratio of the expression values from the treated and untreated library).

be found one would not expect this to happen frequently if the alternative TSS has a distance of only one or two base pairs. In this case this might rather be an artifact of the sequencing procedure or read mapping. For this reason genomic TSS candidates are clustered if they are located on the same strand and are not further than 3 bp apart. From each cluster only the TSS with the highest expression is kept.

6.1.3. Cross-genome comparison of TSS using the SuperGenome

In the next step TSS candidates from different genomes are assigned to each other in order to decide in which genomes a TSS occurs and appears to be enriched. For this the individual genomic TSS are mapped into the SuperGenome. Then the genomic TSS sets are compared and candidates are assigned to each other if they are not further than 1 bp apart on the SuperGenome level. In the next step genomic TSS candidates that have been assigned to each other are summarized to form a SuperGenome TSS (*SuperTSS*). Each SuperTSS carries all information of its genomic TSS including their positions and all properties, i.e., the respective expression height, factor of height change and enrichment factor.

Finally these SuperTSS candidates undergo an additional filtering procedure to increase the confidence of the prediction. For this a stricter set of criteria is applied that have to be fulfilled by at least one genomic TSS of the SuperTSS. In particular the absolute expression height has to be greater than or equal to 0.3, the factor of height change has to be greater than or equal to 2.0 and the enrichment factor has to be 2.0 or above. Thus, a SuperTSS has to be enriched in at least one genome. Genomic TSS are classified as *enriched* if their enrichment factor is at least 2.0 or as *not enriched* otherwise. If the enrichment factor of a genomic TSS is very low, i.e., if the expression value in the untreated library is at least 1.5 times higher than

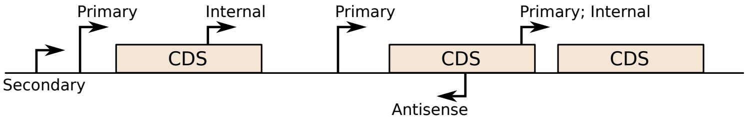


Figure 6.7.: Illustration of TSS classes. TSS can be classified as *Primary*, *Secondary*, *Internal*, *Antisense* or *Orphan* depending on their location relative to annotated genes.

in the treated library, the TSS is considered as a processing site and not classified as detected.

If a SuperTSS was not detected in a certain genome, i.e., if there is no genomic TSS of the respective genome assigned to that SuperTSS, there will still be a TSS candidate annotated if the SuperTSS can be mapped into the respective genome. This is the case if the respective locus of the SuperGenome represents a region of the whole-genome alignment of which the genome was a part of. In such a case the genomic TSS candidate will be classified as *mapped*, but not as *detected*.

6.1.4. Classification of detected TSS

In the next step the final TSS annotations are further classified according to their location relative to annotated genes. This is done for each genome separately. For this a similar classification scheme as previously described by Sharma *et al.* is used [142]. An illustration of the different TSS classes is presented in figure 6.7. For each TSS it is decided if it is the *Primary* or *Secondary* TSS of a gene, if it is an *Internal* TSS, an *Antisense* TSS or if it cannot be assigned to either of these classes. In this case the TSS is classified as an *Orphan*. A TSS is classified as *Primary* or *Secondary* if it is located upstream of a gene's translation start site not further apart than 300 bp. The TSS with the strongest expression is classified as *Primary*. All other TSS that are assigned to the same gene are classified as *Secondary*. *Internal* TSS are located within an annotated gene on the sense strand and *Antisense* TSS are located inside a gene or within a maximal distance of 150 bp on the antisense strand. These assignments are indicated in the respective columns of the MasterTable, as described in detail in the next subsection.

6.1.5. Comprehensive compilation of the results: The MasterTable

The results of the cross-genome TSS detection procedure are primarily presented in the form of a table. Since it contains all important information about the TSS prediction results it is called the MasterTable. It contains for each SuperTSS one row for each genomic TSS that could at least be mapped to its respective genome. If a genomic TSS was assigned to more than one TSS class, there is one row for each classification. A TSS can for example be classified as *Primary* for a certain gene and as *Antisense* for another one.

The MasterTable consists of 28 columns. In the first two columns the position and strand of the SuperTSS are given, which refer to the SuperGenome coordinate system. The next two columns indicate to how many genomes the SuperTSS could be mapped and in how many it was detected. Column 5 contains the identifier of the genome to which the information in the next columns (i.e., the information on the genomic TSS) are referring. The next five columns indicate if the genomic TSS was detected and found to be enriched and they provide detailed values for the expression height, the expression factor and the enrichment factor of the TSS. Column 11 contains the number of classes the TSS was assigned to. The position and strand of the genomic TSS referring to the respective genome are given in the next two columns. The following 5 columns provide information about the annotated gene to which the genomic TSS was assigned during the classification procedure. That is the locus tag of the gene and its functional annotation as well as its length and the length of the 5' UTR in the case of *primary* and *secondary* TSS. Columns 19-22 show the class assignment of the TSS (*Primary*, *Secondary*, *Internal* or *Antisense*). Here, *Orphan* TSS, which are not in the vicinity of an annotated gene, are indicated by zeros in all four columns. The next columns indicate if the TSS was annotated automatically or manually and if it might belong to a novel sRNA/asRNA. All TSS that have been detected by the software are marked as automatically detected. However, the researcher might want to add rows manually, which refer to TSS that have not been detected by the program. This might be the case if the TSS is too weak to exceed the threshold but there are other indications that are not considered by the algorithm. In the last column the upstream sequence of the TSS is provided (50 bp plus the base of the TSS).

The structure of the MasterTable has been designed to allow for very specific manual analyses but also automated processing, e.g., using `R`. Thus, it is very easy to apply customized filters in a spreadsheet application, for example. By this it is possible to apply stricter sets of thresholds. For example a filter can be used that selects only TSS with a certain expression height or only enriched TSS or TSS that were detected in at least 2 genomes or in all genomes. The provided information about each TSS can also be used to understand why a certain prediction or classification has been made. This knowledge can then be used to refine the TSS detection parameters, for example. In addition to that the researcher can define a custom classification scheme and use this for statistical analyses. An example could be the comparison of expression heights of all perfectly conserved UTR TSS with those of all genome specific UTR TSS (classified as *primary* or *secondary*; detected in all genomes vs. detected in one genome).

Furthermore, the additional information on related gene annotations can be utilized to compare UTR lengths or calculated UTR length distributions in general. Additionally, the provided upstream sequence of TSS can be the basis for motif search and promoter analysis. For a detailed use case on how the MasterTable can be exploited in a study see chapter 7, where a transcriptomic study in *Campylobacter jejuni* is presented.

6.1.6. Additional files

In addition to the MasterTable several supplementary result files are generated. All TSS annotations are provided as a GFF file for the SuperGenome coordinate system and for each genome individually. The normalized dRNA-seq graphs are written in wiggle format, again for both coordinate systems. The dRNA-seq graphs in SuperGenome coordinates are especially useful as they can be loaded into a genome browser to verify the predictions manually, which is possible as the expression graphs are aligned via the SuperGenome. Thus, it is in principle also possible to use the software solely for the alignment and normalization of RNA-seq expression data from different genomes. The aligned graphs can then be used for manual analysis or comparative computational methods.

Additionally, all genomic sequences aligned to SuperGenome coordinates are provided as a multi FASTA file, which can also be loaded in a genome browser. Finally, a small table containing some overview statistics is generated. Here, the total number of detected TSS on SuperGenome level and for the individual genomes is provided. These numbers are broken down to show the numbers of TSS in the different classes.

6.1.7. Comparative analysis of different experimental conditions

In general the comparative TSS prediction as described above is designed to be applied to a set of different genomes. However, a different type of study would be the comparison of various experimental conditions applied to the same organism. In this case the SuperGenome transformation that is necessary for the comparison of different genomes can be omitted. Therefore, the software can be run in a cross-condition rather than a cross-genome mode, where no SuperGenome transformation is performed and only one genomic sequence and one genome annotation has to be provided.

Besides this all prediction, comparison and classification procedures are the same, as the SuperGenome, which is used for cross-genome studies, can be seen as an interface between the whole-genome alignment and the applied comparative method. Thus, the methods do not have to be adapted no matter if they are applied to the SuperGenome and therefore a multiple genome alignment or if they are applied to an ordinary genomic coordinate system.

For compatibility reasons the structure of the MasterTable does not differ between the two types of studies. However, some of the columns have a different meaning when experimental conditions are compared instead of genomes. The genome identifier, for example, is replaced by the identifier of the experimental condition. In addition there are still SuperTSS summarizing TSS that have been detected in the different conditions, but in contrast to the genome comparison mode the SuperTSS refer to the genomic coordinate system instead of the SuperGenome.

6.1.8. Generation of cross-genome expression matrices using the SuperGenome

During the course of a comparative TSS prediction the SuperGenome concept not only allows for the cross-genome assignment of TSS but also for the alignment of the genomic dRNA-seq graphs in single-nucleotide precision. Thus, the method can also be utilized for the comparative analysis of gene expression.

For this, the SuperGenome is in a first step used to map individual genomic annotations into its global coordinate system. In the next step genes that are covering the same locus in the SuperGenome are summarized in groups (so-called *SuperGenes*). This is done by starting with the set of annotated genes of one genome and creating a SuperGene for each element. Then the gene annotations of the next genome are compared to the existing SuperGenes. If an annotation covers the same locus as an existing SuperGene, it is added to it. If there is no SuperGene at the same locus, a new SuperGene is created for the annotation. If the annotation is overlapping more than one SuperGene it is added to all of them. In the most complicated situation more than one annotation overlaps the same SuperGene. In this case the respective SuperGene is duplicated for each overlapping annotation and the annotation is added. This procedure is performed for each genome in the study.

Then, for each SuperGene a common interrogatable region is defined that is covered by all genes that are contained in the SuperGene. Thus, if a SuperGene is duplicated during the building process, this results in two SuperGenes that partially contain the same genomic genes. However, their common interrogatable regions will differ as their gene content is not completely the same. Therefore, even the expression values of those genes that the two SuperGenes have in common will not be identical between the SuperGenes.

Finally, the SuperGenome is used to map the interrogatable regions back to the individual genomic coordinate systems. Here, the respective dRNA-seq graphs are used to determine the mean expression value of the region resulting in an expression value for the SuperGene in the respective organism.

The output is an expression matrix with one row for each SuperGene and one column for each experiment. Note that SuperGenes do not necessarily contain a gene for each genome. In this case the missing entry is indicated by a "NA" in the expression matrix. The genes a SuperGene is consisting of are in most cases orthologs. However, as these ortholog groups are defined via the SuperGenome instead of a reciprocal BLAST search, they are not denoted as *orthologous* but as SuperGene. It is also important to note that especially when genes are added to more than one SuperGene or when SuperGenes have to be duplicated, the respective elements are unlikely to be real orthologs. For this reason several thresholds can be set that prevent genes from being assigned to the same SuperGene if they do not overlap significantly or if their pairwise sequence identity is too low.

In addition, it is possible to use an available ortholog mapping as input. In this case the SuperGenes are built according to that mapping instead of using the SuperGenome. The remaining processing steps are the same.

It is important to note that despite of the normalization that is applied to the dRNA-seq graphs the expression matrix cannot considered to be perfectly normalized as the used normalization method only accounts for linear effects, which is sufficient for the TSS detection (see section 6.1.1) but not for comparative expression analysis. Thus, the matrix has to be handled like a raw expression matrix.

6.2. TSSpredator: A user friendly solution for comparative TSS prediction

The generation of the SuperGenome as a representation for a multiple whole-genome alignment and the TSS detection algorithm are technically separate concepts, which have been joined to allow for a comparative analysis across different genomes. However, the initial implementation required a significant amount of manual work and the compilation of a study-specific configuration file, which was also done by hand.

To enable other researchers to use these methods and adapt their application to their needs TSSpredator was developed. TSSpredator is a user friendly framework with an interactive and dynamic graphical user interface that allows for an easy setup of the study. In addition, TSSpredator can be used via its command line interface, which makes the integration into automated pipelines easy.

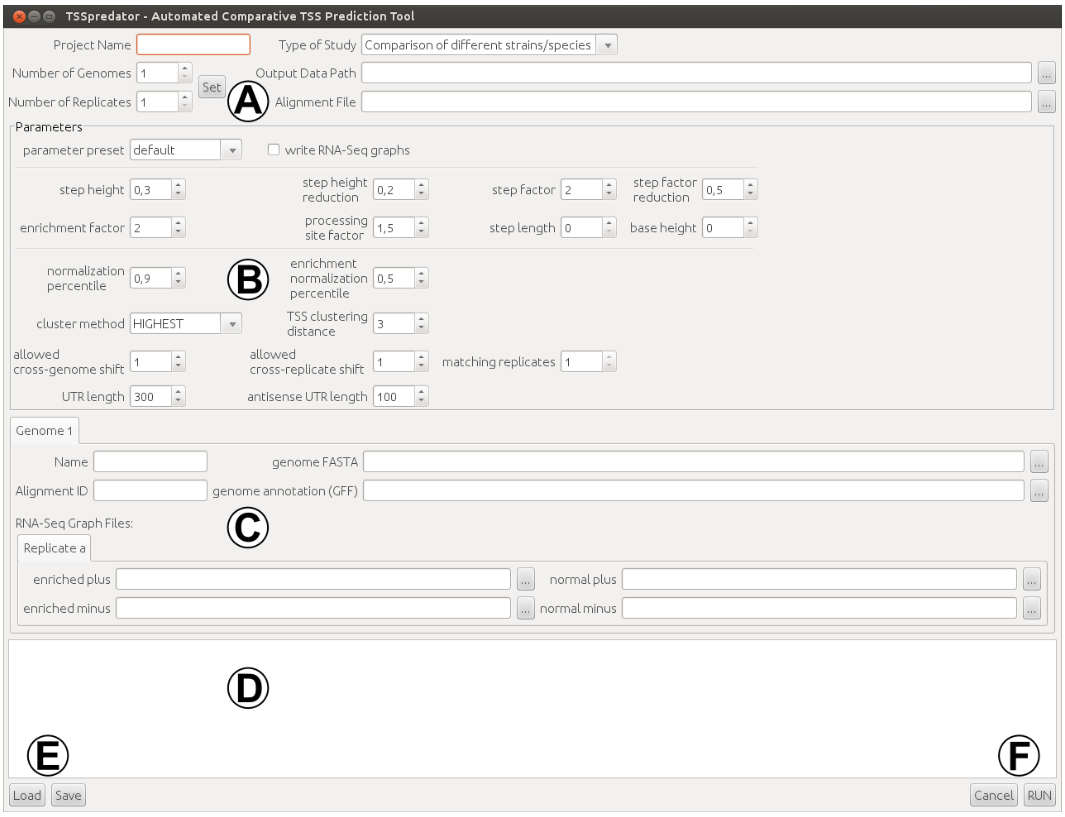
TSSpredator's graphical user interface (GUI) is clearly structured and consists of different areas dealing with different sets of parameters or information that is provided. A screenshot of the GUI is presented in figure 6.8.

In the **study setup area** (Figure 6.8A) general settings for the study can be made. Most importantly these are the type of the study, which is either the comparison of different genomes or the comparison of different experimental conditions, the number of genomes/conditions and replicates, and the path to the output directory. A project name can also be specified.

If the comparative analysis involves different genomes, an alignment file in XMFA format has to be provided. In this case TSSpredator can automatically infer the number of genomes as well as their alignment IDs and names from the XMFA file. The number of replicates has to be set manually. After adjusting the number of genomes/conditions and replicates pressing the *Set* button will create settings tabs for each genome/condition and replicate, respectively.

In the **parameter area** (Figure 6.8B) specific parameters of the TSS prediction procedure can be changed (see section 6.1.2). Instead of changing individual parameters it is also possible to select a parameter preset from the drop-down menu. The presets affect the thresholds for the expression height, the expression factor and the enrichment factor. Currently there are five different parameter presets. Besides the *default* parameters there are two more specific presets (*more specific* and *very specific*) and two more sensitive presets (*more sensitive* and *very sensitive*). With an increased sensitivity TSSpredator detects also TSS of very weakly expressed genes. However, a higher number of false positive predictions has to be expected. When more specific settings are used a higher expression level and a stronger enrichment is required to detect TSS. Thus, the reliability of the predictions is increased but weak TSS will be missed with these settings.

For each genome/condition of the study a tab is generated (Figure 6.8C), in which settings specific for this genome/condition can be made. This includes the name and alignment ID, and file paths to the genomic sequence (FASTA) and the genome annotation (GFF). In addition, for each replicate a tab is displayed within the respective genome/condition tab, where the RNA-seq wiggle file paths of the replicate can be entered.



TSSpredator - Automated Comparative TSS Prediction Tool

Project Name

Type of Study: Comparison of different strains/species

Number of Genomes: 1

Output Data Path

Number of Replicates: 1

Alignment File

Parameters

parameter preset: default

□ write RNA-Seq graphs

step height: 0,3

step height reduction: 0,2

step factor: 2

step factor reduction: 0,5

enrichment factor: 2

processing site factor: 1,5

step length: 0

base height: 0

normalization percentile: 0,9

enrichment normalization percentile: 0,5

cluster method: HIGHEST

TSS clustering distance: 3

allowed cross-genome shift: 1

allowed cross-replicate shift: 1

matching replicates: 1

UTR length: 300

antisense UTR length: 100

Genome 1

Name

genome FASTA

Alignment ID

genome annotation (GFF)

RNA-Seq Graph Files:

Replicate a

enriched plus

normal plus

enriched minus

normal minus

Figure 6.8.: Screenshot of the TSSPREDATOR graphical user interface. A: General settings for the study. B: TSS prediction parameters and other settings. C: Genome/Condition specific settings and files. D: Message area, where information about the prediction procedure is displayed. E: Buttons to *Load* or *Save* a configuration. F: *RUN* button to start the prediction procedure. The *Cancel* button stops a running prediction.

In the **message area** (Figure 6.8D) information about a running prediction process is displayed. Thus, it can be easily determined in which step a running prediction procedure currently is. After the normalization step (see section 6.1.1) the standard normalization factors and the enrichment normalization factors are displayed for each replicate. At the end of the procedure a brief summary is shown indicating the complete number of TSS in the study (SuperTSS) and in the individual genomes/-conditions.

Using the *Save* or *Load* button (Figure 6.8E) a configuration including all settings can be saved, or a previously saved configuration can be loaded, respectively. The configuration is saved in a format that TSSPREDATOR can take as input when running from the command line. Thus, a study can be easily set up using the convenient GUI, the configuration can be saved and used for automated processing in a pipeline.

By pressing the *RUN* button (Figure 6.8F) the prediction procedure is started using the current settings and parameters. A running prediction procedure can be canceled using the *Cancel* button. The first steps of the prediction procedure are the import and normalization of the dRNA-seq data. As this step takes a significant amount of time, the normalized dRNA-seq graphs are cached so that they do not have to be processed again until the program is closed. This allows the user to evaluate different parameter combinations without undergoing the time-consuming data processing every time.

For a study consisting of four different bacterial strains with genome lengths of ∼2 Mb and two biological replicates TSSPREDATOR takes 236 seconds for a complete comparative prediction of TSS including the import of all RNA-seq data and writing the normalized graphs. If the dRNA-seq graphs are already cached and writing of the normalized graphs is switched off, i.e., if the workload is reduced to the TSS prediction, comparative analysis and classification, TSSPREDATOR takes 17 seconds. The memory consumption is below 1 GB. Runtimes have been determined on a desktop PC on one core of an Intel® Core™2 Quad Q9300 (2.5GHz).

7. Novel ncRNAs in the human pathogen Campylobacter jejuni

In 2010, Sharma *et al.* presented a comprehensive study of the primary transcriptome of the major human pathogen *Helicobacter pylori* [142]. For this, a novel differential RNA-seq (dRNA-seq) technique was developed for the enrichment of the 5' end of primary transcripts. The data resulting from this method was utilized for the manual generation of a genome-wide map of transcription start sites (TSS). Using this global TSS map the promoter regions of many annotated genes could be characterized in detail. Furthermore, several candidates for novel non-coding RNAs were detected in addition to an overall high abundance of antisense transcription. This study involved one strain, which was grown under 5 different cultivation conditions in order to increase the depth of the data.

The aim of the work presented in this chapter was the compilation of a global comparative TSS map for four *Campylobacter jejuni* strains [45]. *C. jejuni* is a Gram-negative, microaerophilic pathogenic bacterium that is the major cause of gastroenteritis in human [40, 179]. However, the knowledge about its virulence mechanisms is very limited. Although a type-IV secretion system has been found to be encoded on the plasmid of strain 81-176 [11], which other strains lack, no secretion system could be identified in the genome. It has been hypothesized that other more general capabilities of the bacterium like its motility contribute to its pathogenicity [74, 73, 168]. So far, little is known about the global transcriptomic structure of this organism. Especially the non-coding part of the transcriptome is largely unexplored.

Due to differences between the genomic architectures of the four strains used in this study, mainly insertions, a manual comparative analysis of the data becomes infeasible. Therefore, the SuperGenome concept and the automated TSS detection procedure described in chapter 6 have been applied to the genome-wide detection and comparative characterization of TSS based on dRNA-seq data that was obtained for the four strains of *C. jejuni*. Among other analyses these results were used for the characterization of promoter regions and the identification of novel sRNAs in *C. jejuni*.

7.1. Comparative prediction of TSS in four C. jejuni strains

The basis for the comparative TSS prediction were differential RNA sequencing data produced for each of the four *C. jejuni* strains of the study. These strains are *C. jejuni* RM1221, NCTC11168, 81-176 and 81116. Strain RM1221 was isolated from chicken, the other strains were isolated from human. The size of the chromosomes of these strains varies around 1.6 Mb except for strain RM1221, whose chromosome has a size of about 1.8 Mb, which is mainly due to four large insertions, so-called *Campylobacter jejuni*-integrated elements (CJIEs) [47, 116], most of which are integrated prophages. This also accounts for its higher number of annotated open reading frames, which is 1, 838 for RM1221, where the other genomes have about 1, 600 annotated open reading frames. The GC-content of the four genomes is about 30%.

The bacterial strains were grown under microaerobic conditions and samples were taken from two biological replicates at mid-exponential growth phase. For each replicate two libraries were constructed. For one library RNA was treated with 5'-phosphate-dependent terminator exonuclease in order to deplete processed transcripts as described by Sharma *et al.* [142]. The other library was not treated. The libraries were sequenced using an Illumina HiSeq 2000 in single read mode.

Reads have been mapped against the respective reference genomes using `segemehl` [72]. dRNA-seq graphs containing the number of covering reads for each genomic position were generated using the *Integrated Genome Browser* [107]. For each library two graphs were produced, one for the forward strand and one for the reverse strand. Each graph was then normalized by the total number of mapped reads. These graphs were then used as input for the comparative TSS prediction procedure, where they were further normalized and processed as described in chapter 6.

For the generation of the SuperGenome (chapter 5) a multiple whole-genome alignment of the four strains was generated using the `progressiveMauve` algorithm [38] of the genome alignment software `Mauve` [37]. In order to improve the sensitivity of the alignment the *seed families* option of the aligner was set, which allows for a limited number of mismatches during the seed matching process. Besides that standard parameters were used.

The resulting SuperGenome has a length of 2, 115, 274 bp, of which 1, 380, 020 alignment columns (∼65%) are perfectly conserved among the strains.

The TSS prediction was performed using default parameters. On the genome level a TSS candidate was required to be detected in both replicates to be accepted for the respective genome.

Overall 3377 TSS were detected on the SuperGenome level. An overview of the different numbers of conserved and specific TSS in the different strains or classes is presented in table 7.1. On the genome level these represent between 1905 TSS in strain NCTC11168 and 2167 TSS in strain RM1221. That RM1221 shows the highest number of detected TSS is not surprising because of its large genomic insertions. The number of TSS that are conserved in all four strains is 1035. A comparable number of 1067 TSS are conserved in two or three strains. Altogether 1275 TSS were found to be specific for a single strain. In the individual strains the number of specific TSS varies between 246 in strain NCTC11168 and 450 in strain RM1221, which, again, is due to the larger genome of this strain.

The conservation of TSS in the four strains is illustrated in figure 7.1A. Figure 7.1C shows the overlap of TSS sets if the TSS are not required to be detected in both replicates. It turns out that between 82% and 88% of all TSS in the final set are confirmed by both replicates. The enrichment rates are also quite similar when comparing the strains. Figure 7.1B shows the respective Venn diagram if only enriched

Table 7.1.: Number of detected TSS in the different TSS classes.

	All	Primary	Secondary	Internal	Antisense	Orphan
Number of SuperTSS:	3377	973 (29%)	327 (10%)	1217 (36%)	1624 (48%)	61 (2%)
Conserved in all strains:	1035	527 (51%)	92 (9%)	328 (32%)	445 (43%)	10 (1%)
Conserved in 2 or 3 strains:	1067	204 (19%)	118 (11%)	398 (37%)	534 (50%)	36 (3%)
TSS in NCTC11168:	1905	675 (35%)	180 (9%)	653 (34%)	843 (44%)	17 (1%)
TSS in 81-176:	2003	676 (34%)	193 (10%)	651 (33%)	937 (47%)	20 (1%)
TSS in 81116:	1944	690 (35%)	181 (9%)	653 (34%)	856 (44%)	25 (1%)
TSS in RM1221:	2167	744 (34%)	202 (9%)	735 (34%)	988 (46%)	25 (1%)
Specific for NCTC11168:	246	46 (19%)	19 (8%)	106 (43%)	119 (48%)	3 (1%)
Specific for 81-176:	260	34 (13%)	24 (9%)	98 (38%)	141 (54%)	4 (2%)
Specific for 81116:	319	58 (18%)	35 (11%)	128 (40%)	150 (47%)	5 (2%)
Specific for RM1221:	450	104 (23%)	39 (9%)	159 (35%)	235 (52%)	3 (1%)
Only missing in NCTC11168:	109	22 (20%)	11 (10%)	36 (33%)	59 (54%)	4 (4%)
Only missing in 81-176:	97	25 (26%)	8 (8%)	38 (39%)	40 (41%)	4 (4%)
Only missing in 81116:	165	33 (20%)	24 (15%)	61 (37%)	83 (50%)	6 (4%)
Only missing in RM1221:	99	21 (21%)	10 (10%)	35 (35%)	50 (51%)	3 (3%)

Number of detected TSS in the different TSS classes with respect to the strains the TSS were detected in. The table shows all detected TSS regardless of their enrichment. The percentages are related to the respective number of TSS summed over all classes.

TSS are considered. The enrichment rates vary between 90% and 93%. This small variation indicates that the RNA-seq graph were sufficiently normalized with respect to the enrichment strength. However, evaluation of enrichment rates for conserved TSS shows that only 79% of the TSS that are conserved in all four strains are also enriched in all strains. These elements should still be considered as TSS of primary transcripts as it seems to be quite unlikely that they represent primary transcripts in some strains and processing sites in other. Nevertheless, this possibility cannot be excluded. Therefore, the TSS detection algorithm reports all elements that are found to be enriched in at least one strain. As information about the enrichment is included in the MasterTable, specific filtering can be applied during postprocessing steps in the case that only TSS enriched in all strains are of interest. In this study, however, all TSS enriched in one strain or more were considered for the final global TSS map.

Although the overall enrichment rates do not differ significantly between the strains, a more detailed evaluation of the individual enrichment factors of conserved TSS shows only a correlation of about 0.5 between each pair of strains, whereas the correlation of the respective TSS expression heights varies between 0.69 and 0.78.

The number of TSS in the different classes (*Primary*, *Secondary*, *Internal*, *Antisense* or *Orphan*) is significantly influenced by the conservation of the TSS. Regarding TSS conserved in all four strains about 60% are classified as UTR TSS (*Primary* or *Secondary*). Among the TSS that are specific for an individual strain only between 22% and 32% are classified as UTR TSS, while the respective fraction of Internal and Antisense TSS is higher. The proportion of *Orphan* TSS is also rising to about 4% in contrast to about 1% when considering conserved TSS. This picture indicates that UTR TSS are significantly more likely to be conserved than other classes and that strain specific TSS tend to have a higher fraction of TSS that are *Internal*, *Antisense* or *Orphan*. However, in general the fraction of *Internal* and especially *Antisense* TSS is quite high. On the SuperGenome level 1624 TSS were detected that were classified as *Antisense*. In the individual strains they represent between 843 and 988 *Antisense* TSS, which corresponds to about 45% of the TSS detected in the respective strains. Besides this immense amount of antisense transcription the fraction of *Internal* TSS is also quite high constituting about 33% of all TSS detected in a strain. While the antisense transcripts might have a regulatory function the role of internal sense transcripts is even more unclear.

As the TSS classes are not disjoint a TSS can be assigned to more than one class. The overlap of TSS classes in the different strains is illustrated in figure 7.2. About 25% of all *Primary* TSS are also classified as *Internal* and about the same fraction is also classified as *Antisense*. In each strain between 20 and 30 TSS are classified as *Internal* but are also located *Antisense* to a different gene. This shows that a significant amount of antisense transcription and also internal transcription starts is actually due to protein-coding transcripts with overlapping genomic location. However, about 75% of all *Antisense* or *Internal* TSS cannot be explained by this and therefore potentially represent independent transcripts.

7.2. SNP–dependent strain-specific promoter usage

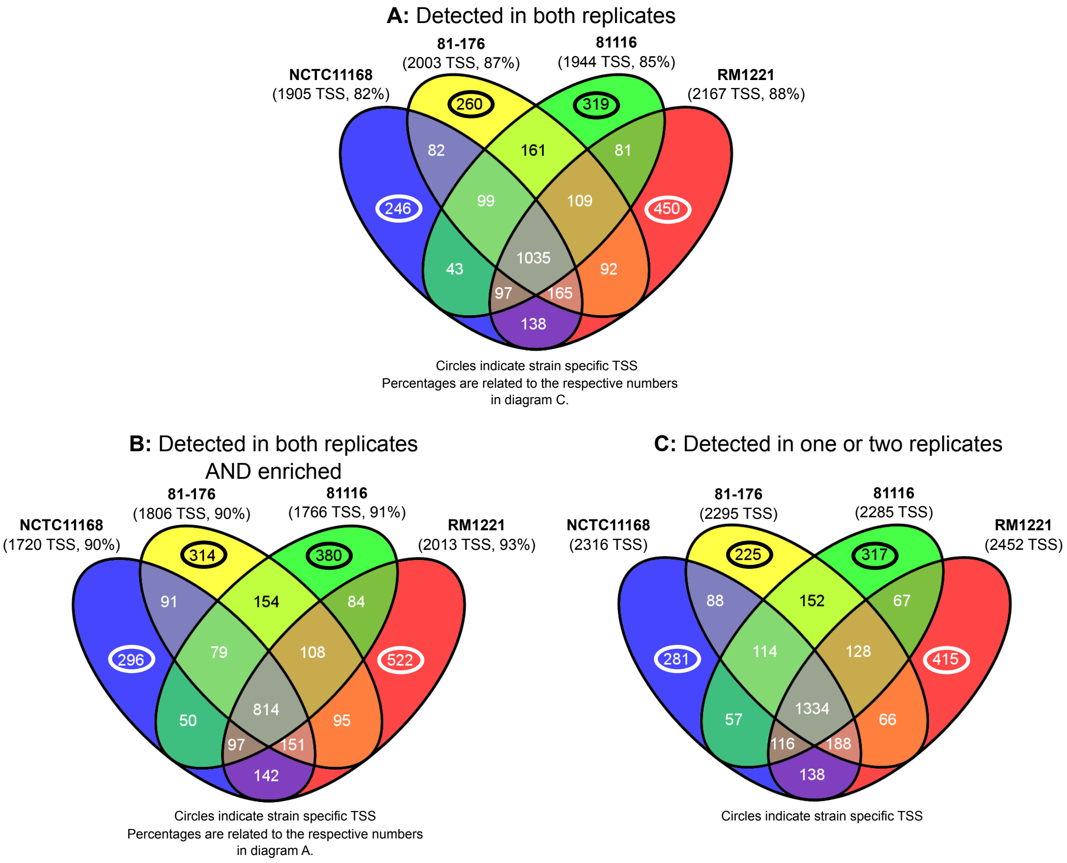
The comparative analysis of the global TSS maps of the four *C. jejuni* strains allowed for the identification of promoters with a high level of sequence conservation, but their downstream genes show extremely different expression levels. It turned out that in many of these cases single nucleotide polymorphisms (SNP) can lead to a strain-specific promoter usage. I.e., the detection of SNPs in the respective promoter regions of the four strains correlates with the expression that is observed for the corresponding TSS.

An example for this is an internal TSS of the *pnk* gene as illustrated in figure 7.3. The promoter region of the primary TSS of this gene is perfectly conserved among the four strains and a strong expression is observed in all of them. However, the internal TSS, which is located further downstream, is only expressed in strains RM1221 and NCTC11168. The respective promoter in the other two strains, 81-176 and 81116, seems to be disrupted by a SNP in the -10 box.

Other examples of strain-specific promoter usage revealed that mutations in the A/T-rich region upstream of the -10 box might also be responsible for a loss of transcription. This indicates that this region is either directly or indirectly relevant during transcription initiation.

7.3. Novel ncRNAs in C. jejuni

On the basis of the automated cross-genome TSS prediction it was possible to identify TSS that potentially belong to novel non-coding RNAs in addition to those that could be assigned to protein-coding genes. The fraction of antisense TSS in each



A: Detected in both replicates

Circles indicate strain specific TSS
Percentages are related to the respective numbers
in diagram C.

B: Detected in both replicates
AND enriched

Circles indicate strain specific TSS
Percentages are related to the respective numbers
in diagram A.

C: Detected in one or two replicates

Circles indicate strain specific TSS

Figure 7.1.: **A:** Venn diagram showing the overlap of the TSS sets detected in both replicates of the respective strains. The percentages next to the indicated numbers of TSS are related to the respective TSS set that was detected in at least one replicate. The fraction of TSS detected in both replicates varies between 82% and 88%. **B:** Venn diagrams showing the overlap of the TSS sets detected in both replicates and found to be enriched in the respective strains. The percentages next to the indicated numbers of TSS are related to the respective numbers in diagram A. The fraction of enriched TSS varies between 90% and 93%. **C:** Venn diagram showing the overlap of the TSS sets detected in at least one replicate of the respective strains.

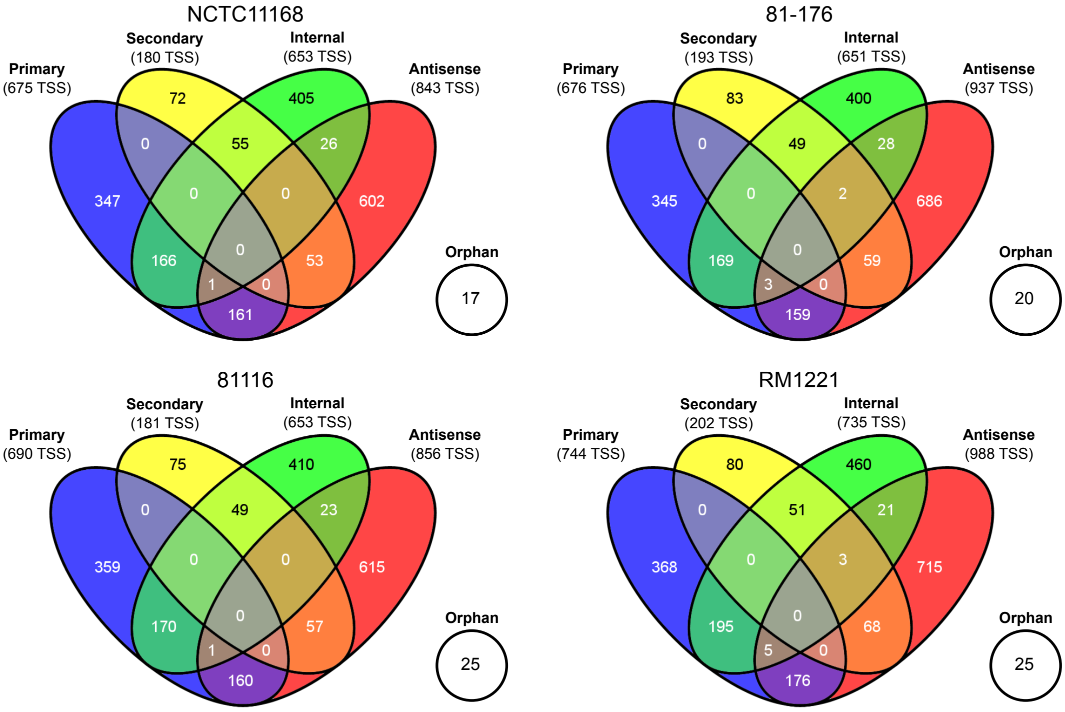


Figure 7.2.: TSS classifications for the individual strains. The Venn diagrams indicate the overlap between TSS classes for the four individual strains. Many TSS are assigned to more than one class. For example in NCTC11168, 167 of the 675 *Primary* TSS (∼25%) are also classified as *Internal* and 162 (∼24%) are also classified as *Antisense*. In 81-176, 172 of the 676 *Primary* TSS (∼25%) are also classified as *Internal* and 162 (∼24%) are also classified as *Antisense*. In 81116, 171 of the 690 *Primary* TSS (∼25%) are also classified as *Internal* and 161 (∼23%) are also classified as *Antisense*. In RM1221, 200 of the 744 *Primary* TSS (∼27%) are also classified as *Internal* and 181 (∼24%) are also classified as *Antisense*. Venn diagrams were generated by VENNY (http://bioinfogp.cnb.csic.es/tools/venny/index.html).

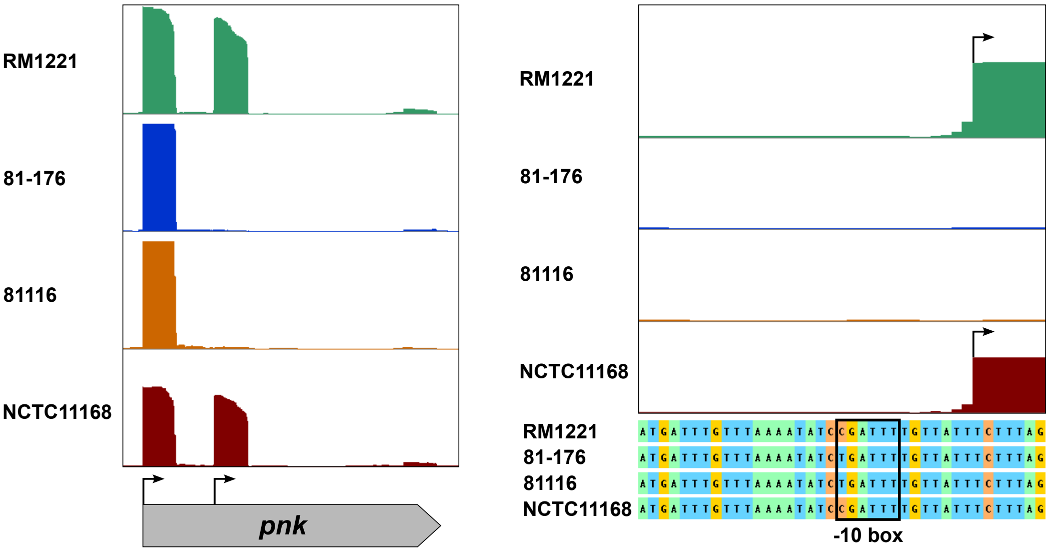


Figure 7.3.: Example of an internal TSS in *C. jejuni*, whose expression is affected by a mutation in its promoter sequence. Left: RNA-seq data shown for the genomic locus of the *pnk* gene. The RNA-seq expression graphs of the four *C. jejuni* strains RM1221, 81-176, 81116 and NCTC11168 have been mapped to the SuperGenome coordinate system for a direct visual comparison. The primary TSS of the *pnk* gene, which coincides with the translation start, is expressed in all four strains. Transcription of the internal TSS, which is located further downstream, is only observed in the two strains RM1221 and NCTC11168. Right: Close-up view of the promoter region of the internal TSS showing RNA-seq data and sequence information in the SuperGenome coordinate system. In strains 81-176 and 81116, where no transcription was measured at the internal TSS, a mutation from C to T compared to the other strains can be observed within the -10 box of the promoter.

genome is about 45%. The high abundance of antisense TSS suggests that there is potentially a large repertoire of *cis*-encoded antisense RNAs in *C. jejuni*. In addition, several candidates for *trans*-encoded ncRNAs could be identified. The expression of most candidate ncRNAs could be verified by Northern blots. BLAST analyses revealed that on the sequence level many ncRNAs are only conserved in *C. jejuni*. Few elements, however, also show conservation in other *Campylobacter* species or even in *Helicobacter*. In contrast some other elements are not even conserved among the four *C. jejuni* strains of the study. These elements might therefore be involved in the regulation of strain-specific mechanisms. Furthermore, some of the ncRNAs conserved in all four strains were not found to be expressed in all of them.

The automatically detected TSS also helped to characterize a type-II CRISPR system (clustered regularly interspaced short palindromic repeats) in three of the four investigated strains (see also section 2.1.5). In strain 81-176 the CRISPR locus cannot be found and only a weak expression of the system is detected in strain RM1221, where the *cas9* gene, which is required for crRNA stabilization carries a stop mutation. In type-II CRISPR systems a ncRNA termed TracrRNA is involved in crRNA maturation [41, 98]. From the dRNA-seq analysis it is evident that this RNA as well as the crRNAs are transcribed. Interestingly, enriched TSS were found upstream of each repeat-spacer unit. Thus, in the investigated *C. jejuni* strains each crRNA unit carries its own promoter, whereas in other similar systems that have been characterized a single transcript is produced, which consists of all crRNAs and an upstream leader sequence.

8. A pipeline for the genomic reconstruction and analysis of ancient pathogens

As demonstrated in chapter 7 the comparative integration of transcriptomic and genomic data can reveal small differences in the genome, like single nucleotide polymorphisms (SNP), which can alter the transcriptomic output of an organism significantly. Therefore, such analyses can help to identify the causes of certain phenotypic differences, for example in the context of pathogenicity.

Understanding why specific bacterial strains are virulent for humans or life stock while other strains of the same species are not is of major importance for modern medicine as this knowledge can lead to the development of new treatments.

However, studying these differences can also be of historical interest. In the field of paleogenetics, ancient DNA (aDNA) sampled from old remains such as bones is sequenced and used to reconstruct the genomes of ancient organisms, which can be the human genome or the genomes of ancient pathogens [171]. One important question in this context is why infectious diseases like the plague or leprosy caused such devastating pandemics in the medieval time while they seem to be much less virulent nowadays.

In 2011, Bos *et al.* were able to isolate DNA of a medieval strain of *Yersinia pestis*, which was responsible for the Black Death, claiming about 100 million victims worldwide during the medieval time [24, 140]. Using an array enrichment technique it was possible to sequence enough *Y. pestis* DNA to generate a draft genome that could be compared to modern reference genomes. This comparison showed no unique SNPs in the medieval strain indicating that the decreased virulence of modern strains cannot be explained by genomic variation alone.

To conduct these comparative analyses between ancient bacterial strains and modern strains in a systematical and reproducible manner, experimental as well as computational methods are needed that consider the specific properties of ancient samples [124]. One of the issues that have to be dealt with is the short length of the aDNA fragments. This makes classical paired-end sequencing difficult or even impossible. The short fragments can still be sequenced from both ends, but the two reads will probably overlap at their 3' ends. Thus, their is no real mate-pair information that can be exploited during read mapping or *de novo* assembly. However, the fact that the two paired reads are overlapping can be used to correct for sequencing errors in the overlapping region. For this, the two overlapping paired-end reads reads are merged into one single read. Any discrepancies regarding the two sequences of the overlapping region can be solved by taking the sequence information from the read that has a higher base calling score at the respective position. An additional advantage of this read merging procedure is an increased length of the resulting single reads, which leads to a higher confidence in mapping. Because of the usually low genomic coverage in aDNA projects it is important to achieve high mapping confidence value in order to keep as many reads a possible for the SNP analysis. Reads with a low mapping confidence have to be discarded as they could lead to false positive SNP calls.

Another challenge in aDNA processing is the occurrence of DNA damage [27]. These damage patterns can be used to determine if the sampled DNA is really of ancient origin and the amount of damage can in some cases even be used to roughly estimate the age of the DNA [56]. However, for further processing the DNA is usually chemically repaired [28], in order to decrease the noise during SNP detection, for example. For repaired aDNA fragments no damage pattern has to be taken into account during the analysis of the sequencing reads.

The fact that DNA fragments are shorter in ancient samples and that mate pairs are not available also affects *de novo* assembly. To assemble ancient genomes without relying on a reference sequence can be important to assess variations of the genomic architectures of ancient and modern genomes, which cannot be done by read mapping alone. In addition to the shorter fragment length in general, the reads resulting from the sequencing are of different length after postprocessing, which is due to the merging of read pairs and lower base call scores at the 3' end, which often makes a trimming of the reads necessary. Most short read assemblers such as `SOAPdenovo` [90, 96], `Velvet` [181] or `ABySS` [143] (see [134, 25] for reviews) are based on a de Bruijn graph that indirectly models the overlap between reads by splitting them into k-mers, which form the vertices of the graph. Overlaps between k-mers are represented by directed edges. Contigs are constructed by the identification of Eulerian or Hamiltonian paths in the graph. The topology of the graph and the performance of the approach are highly dependent on the chosen k-mer size. The optimal value for k depends on various factors such as the read length, the error rate and the number of reads in relation to the expected genome size. If the chosen value is too small, bridging even small repeat regions is not possible anymore. In addition, a small k leads to a graph with a large number of vertices, which is very complex and more difficult to process. Larger values for k, on the other hand, make the detection of small overlaps between reads impossible. This is especially relevant if the reads are short and the estimated genomic coverage is low. Furthermore, reads that are shorter than the selected k-mer size cannot be considered during the assembly. However, if the reads vary in length it might not be possible to avoid the exclusion of very short reads.

Most of the mentioned factors can be problematic in the context of ancient genome assembly. The reads are of varying length and often quite short and the total amount of DNA that originates from the target organism is low, which leads to a low genomic coverage. The fact that ancient samples actually are metagenomic samples in all cases is also one of the biggest challenges in the context of *de novo* assembly. For these reasons it is usually not possible to estimate the optimal k-mer size prior to the analysis and it is not sufficient to only test a small range of values. Instead, an exhaustive search for the optimal parameter is necessary.

A proper *de novo* assembly of an ancient genome is not possible in many cases due to a low amount of endogenous DNA and, therefore, a low genomic coverage. However, a partial reconstruction based on read mapping to a reference sequence might still be possible. For multiple samples this reconstruction has to be conducted in an efficient and reproducible manner.

Computational tools for the specific preprocessing and analysis of aDNA sequencing data have been developed earlier [83]. In addition, for the mapping of aDNA sequencing reads to reference sequences the software `MIA` is available, which is able to consider specific properties of aDNA, such as damage patterns, during the mapping process [58]. `MIA`, however, has been developed for rather small reference sequences such as the mitochondrial genome, but due to its runtime and memory consumption it is not applicable to full prokaryotic or eukaryotic genomes. Thus, for the analysis of ancient bacterial genomes fast read mappers have to be applied, which are primarily designed to work on sequencing data of modern DNA. Therefore, the adaptation of various parameters to the properties of aDNA is potentially necessary [139]. In addition, tools for comparative genomic analyses have to be applied in order to elucidate variation between ancient and modern genomes.

In this chapter I will present an automated computational pipeline for the reconstruction and comparative analysis of ancient genomes. This pipeline has been developed and applied in the context of the genome-wide comparison of medieval and modern strains of *Mycobacterium leprae*, a bacterial pathogen causing leprosy [141]. The details of each step are described in the following sections.

8.1. A pipeline for ancient bacterial genome reconstruction and analysis

The pipeline presented here addresses all steps starting from the preprocessing of sequence data to read mapping and comparative SNP calling. It consists both of state of the art tools as well as self developed scripts. A schematic representation of the workflow is depicted in figure 8.1.

After merging of overlapping read pairs the quality of the resulting reads is assessed and bases of low sequencing quality are trimmed from the 3' end. The remaining reads are used for *de novo* assembly or mapping to a reference genome. SNPs are called on the basis of the mapping and after filtering the SNPs are used to generate a draft genome sequence for each sample. The draft genomes of all samples are aligned in order to call SNPs comparatively in all samples. The alignment of SNP positions is used to reconstruct the phylogeny of all samples. Furthermore, the effect of the SNPs on protein-coding genes is predicted.

8.1.1. Preprocessing and Mapping

In the first step the sequencing quality of all samples is assessed using `FastQC` [8]. Adapters are removed and by default nucleotides with a phred score smaller than 20 are trimmed from the 3' end. All reads with a length of 30 nucleotides or longer are

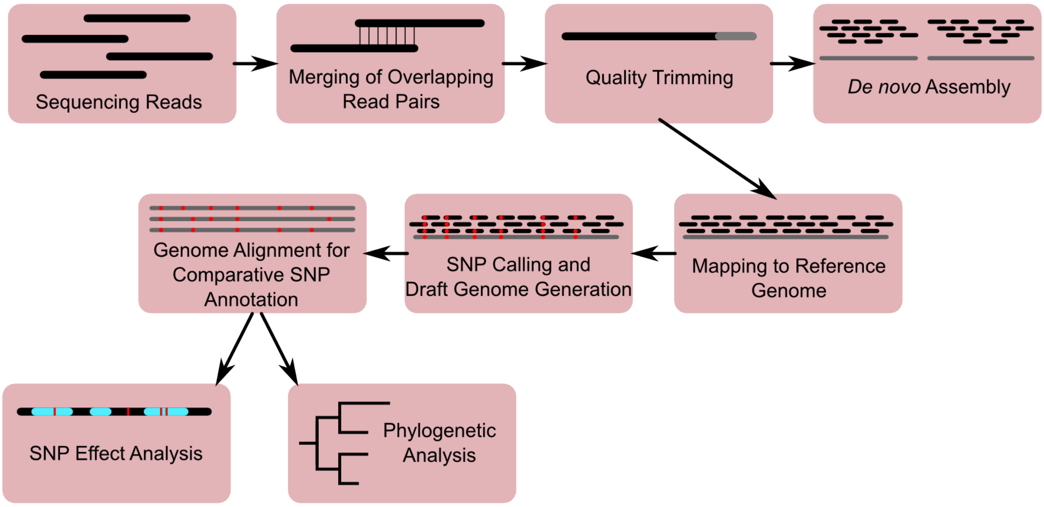

Sequencing Reads
Merging of Overlapping Read Pairs
Quality Trimming
De novo Assembly
Mapping to Reference Genome
SNP Calling and Draft Genome Generation
Genome Alignment for Comparative SNP Annotation
SNP Effect Analysis
Phylogenetic Analysis

Figure 8.1.: Schematic representation of the aDNA processing pipeline. Overlapping read pairs are merged and trimmed to preserve only bases with high sequencing quality. The reads are then used as input for a *de novo* assembly or for mapping to a reference genome. Based on the reference mapping, SNPs are detected and used to generate draft genome sequences for all samples, which are then aligned in order to detect SNPs comparatively in the complete data set. Aligned SNP position are used for phylogenetic analyses. In addition, the effects of SNPs on protein-coding genes are determined.

kept. For samples that are subject to paired-end sequencing with overlapping reads, which usually applies to the ancient samples, overlapping read pairs are merged, if the size of the overlap region is at least 10 nucleotides.

The remaining reads of all samples are then mapped to a common reference genome using the Burrows-Wheeler Aligner (`BWA`) [89]. For this the `BWA` *samse* algorithm is applied.

8.1.2. Mapping assemblies

The next steps are concerned with the genome reconstruction. For this, the read mappings of all sequenced samples are used to create a so-called mapping assembly for each strain. In comparison to a reference-free *de novo* assembly the draft sequence is generated by evaluating the results of the mapping as described in the following. In a first step, the `UnifiedGenotyper` module [42] of the *Genome Analysis Toolkit* (GATK) [101] is applied to each mapping, i.e., to each `bam` file. GATK produces a `vcf` file (variant call format [36]) as output. The `EMIT_ALL_SITES` flag of `UnifiedGenotyper` is set to make sure that the resulting `vcf` files contain one entry for each genomic position of the reference genome. Thus, reference bases as well a variant positions are called and emitted.

In the second step the Java tool `VCF2Draft`, which was developed during this dissertation, is applied to the `vcf` file of each sample to generate the draft sequences.

VCF2Draft reads a vcf file row by row and incorporates for each row and thus for each call that was made by GATK one nucleotide into the new draft sequence. Using default parameters it incorporates a reference base, if the quality of the respective call was at least 30 and the position was covered by at least 5 reads. A variant call (SNP) is incorporated, if the same quality threshold is fulfilled, if at least 5 reads covering the respective locus contain the SNP and if the fraction of mapped reads containing the SNP was at least 90%. If not all of these requirements for a variant call are fulfilled, but the quality threshold is still reached, the reference base is called instead, but only if it is confirmed by at least 5 reads. If neither a reference call nor a variant call can be made, the character 'N' is incorporated at the respective position.

In addition to the draft sequence that contains 'N's at positions without a specific call, two further draft sequences are generated. First, a draft sequence that contains the reference base instead of 'N' is generated. This sequence is used during the multiple whole-genome alignment step as the alignment quality is much higher when avoiding ambiguous bases (see 8.1.3). Second, a draft sequence with a special uncertainty encoding is generated. Instead of the 'N' character it contains the numbers 1, 2, 3, 4 encoding A, C, G, T at positions where a call was rejected (e.g. due to low coverage) but the reads covering the respective position unambiguously indicate a certain nucleotide call. Using this special draft sequence allows for the differentiation between a clear SNP call, a weak SNP call, a clear/weak reference call and no call ('N') at a certain position.

8.1.3. MAUVE Alignment

To compare variant positions among strains including multiple published reference genomes, a multiple whole-genome alignment of all generated draft sequences and the genomic sequences of selected reference sequences is computed using the progressiveMauve algorithm [38] integrated in the whole-genome alignment software Mauve [37]. The SNP export function available in Mauve is used to generate a list of all alignment columns in which at least one strain contains a SNP in comparison to the primary reference sequence, which was used for mapping.

This list is then subject to further processing, for which the Java program SNPtableAnalyzer was developed in this dissertation. In a first postprocessing step, those alignment columns are labeled that represent SNPs that are located in regions that should be excluded from the analysis. This allows for their exclusion during phylogenetic analysis. In the final SNP table that is generated by the pipeline these SNPs are listed but indicated as excluded from further analysis. Regions that are excluded from the analysis can be repeat regions or loci to which metagenomic reads from other organisms are mapping. They can be identified by an environmental, negative control sample that does not contain the target organism.

The concatenation of the filtered alignment columns is then converted into fasta format and can be used as input for MEGA5 [151] or BEAST [44] for example to perform the phylogenetic and the dating analysis, respectively.

In addition to the output for phylogenetic analyses SNPtableAnalyzer produces an input file for SnpEff in order to predict effects of SNPs on protein-coding genes.

8.1.4. SNP effect analysis

For further analyses with respect to the effect of SNPs on annotated genes the software `SnpEff` [33] is applied to the identified variant positions. If non-coding genes and pseudogenes should be considered during the `SnpEff` analysis, a custom database has to be constructed as these elements are not part of the databases that are provided by `SnpEff`. The `SnpEff` parameter for the up-/downstream region size for reporting SNPs that are located upstream or downstream of genes is set to 100 nt. For all other parameters default values are used.

`SNPtableAnalyzer` is applied to the `SnpEff` output and the original `Mauve` output simultaneously to compile a comprehensive table providing information on each SNP regarding its effect on annotated genes and the strains in which the SNP occurs. For this `SNPtableAnalyzer` utilizes the draft sequences produced with the special uncertainty encoding to distinguish clearly called SNPs and reference bases from positions, where the thresholds for calling were not reached due to low coverage, for example, but where the reads mapping to the respective locus still indicate either a SNP or a reference call. This is to improve the overall interpretability of the results presented in the final SNP table.

8.2. Application to the comparison of medieval and modern Mycobacterium leprae strains

The analysis pipeline was applied to the comparative analysis of 5 leprosy samples from different skeletal remains (Jorgen_625, Refshale_16, 3077, SK8, SK2), 7 samples from recent biopsies of leprosy patients (S2, S9, S10, S11, S13, S14, S15) and 4 reference strains (TN, Br4923, Thai53, NHDP63). In the samples the fraction of leprae DNA was increased using an array enrichment technique. In the Jorgen_625 sample the amount of endogenous DNA was so high (40%) that an enrichment step was not necessary. Medieval and modern samples were then subject to high-throughput sequencing and draft genomes were generated by mapping the sequencing reads to *M. leprae* TN as a global reference. Further analyses steps included the comparative annotation of SNPs among the draft genomes of the samples and fully sequenced reference strains. In addition, the well preserved DNA in the Jorgen_625 sample allowed for a *de novo* assembly and thus the unbiased and reference-free generation of a draft genome that could be used to search for potential changes in genome architectures between medieval and modern strains.

To elucidate the origin and the development of *Mycobacterium leprae* the identified variant positions were used for phylogenetic analyses.

8.2.1. Mapping assemblies and SNP analysis

As the common reference for read mapping the genome of *Mycobacterium leprae* TN was chosen. Considering the thresholds for reference base and SNP calling (coverage $\geq$ 5 reads; quality $\geq$ 30; major allele frequency $\geq$ 90%) the fraction of the genome that could be reconstructed was between 83.8% (sample 3077) and 98.3% (sample

Jorgen_625) for the medieval samples and between 84.4% (sample S2) and 97.6% (sample S11) for the modern samples. Thus, the values for the genomic coverages do not differ significantly between medieval and modern samples and the sample with the highest coverage is Jorgen_625, for which the array enrichment was not necessary. This indicates a surprisingly high preservation of *M. leprae* DNA in the medieval samples.

In the draft sequence generation procedure the number of called variant positions ranged from 62 (sample 3077) to 115 (sample SK2) for the medieval samples and from 65 (sample S2) to 217 (sample S15) for the modern samples. Again, there is no significant difference with respect to the number of SNP calls between medieval and modern genomes except sample S15, which has an extraordinary high number of variant calls, which might be due to higher selection pressure resulting from anti-leprosy treatment [97].

The alignment of the reconstructed draft genomes of the 5 medieval strains and the 7 modern strains with the published reference genomes *M. leprae* TN, Br4923, Thai53 and NHDP63 led to the identification of altogether 755 variant position, of which 723 remain after the exclusion of regions to which reads from the negative control sample SK12 mapped. This sample was taken from a skeleton of the same medieval cemetery as sample SK2 with no indication of an infection with *M. leprae*. For the SNP effect analysis a custom gene annotation database for the reference *M. leprae* TN was built using the respective GFF file from NCBI (NC_002677.gff). The analysis revealed that 349 of 723 SNPs cause non-synonymous changes in annotated genes. 141 of those 349 SNPs are located in pseudogenes. Thus 208 SNPs affect genes producing a potentially functional protein. The gene with the highest number of non-synonymous SNPs is the cell surface protein ML0411. It contains 10 non-synonymous SNPs of which one was detected in all strains (sample SK14 has insufficient coverage at that position). One SNP was specifically detected in the medieval sample Refshale_16. Each of the other 8 SNPs were found in one or two modern samples.

Overall, the genomic diversity between ancient and modern strains is quite low. Therefore, the decline of leprosy in Europe was probably not caused by a reduced virulence but by other factors like host immunity or improved hygiene conditions.

8.2.2. Phylogenetic analysis

On the basis of the 723 variant positions between the 5 medieval strains, the 7 modern strains and the 4 reference genomes a phylogenetic analysis was performed. For this, the genome of *Mycobacterium avium* 104 (NC_008595.1) was included as an outgroup. However, the set of variant positions was restricted to sites that show variation among the *M. leprae* strains. The rooted phylogenetic tree was inferred using maximum parsimony as implemented in `MEGA5` [151]. Sites with more than 10% alignment gaps or missing data were excluded, which restricted the data set to 537 sites. Bootstrap values were computed from 500 repetitions. The resulting phylogenetic tree is depicted in figure 8.2. The five medieval samples form two distinct branches. Samples Jorgen 625 and SK2 cluster with the modern strain NHDP63. The

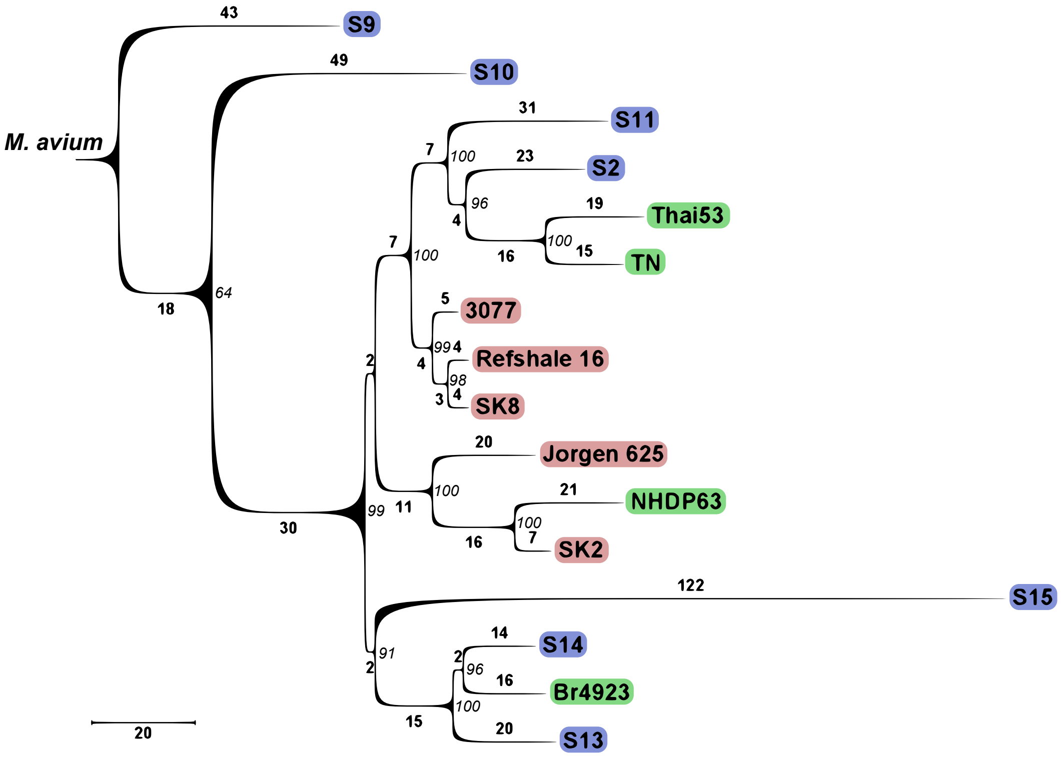


Figure 8.2.: Maximum Parsimony tree of ancient and modern *M. leprae* strains. Ancient strains are highlighted in red, modern strains sequenced in this study are highlighted in blue and publicly available full genomes are highlighted in green. Internal nodes are labeled with bootstrap statistics. Branches are labeled with branch lengths representing the absolute number of substitutions.

other branch consists of the medieval strains 3077, Refshale 16 and SK8. Notably, none of the modern strains falls on this branch. Sequencing of more modern *M. leprae* strains will help to investigate if these three medieval strains might represent an extinct lineage.

8.2.3. De novo assembly

Due to the well preserved *M. leprae* DNA in the Jorgen_625 sample it was possible to perform a *de novo* assembly of the respective metagenome. For this the short read assembly software `SOAPdenovo` [90] was used. Several different values for the k-mer size parameter were evaluated. All possible k-mers from k=107 to k=127 were evaluated with k=127 yielding the best results. Furthermore, three smaller values for k were also tested (k=49, k=79, k=93) in order to check if there is a second optimum regarding assembly performance, which was not the case. As quality measures the number of contigs, the average and maximal contig size, the contig N50/N90 and the total length of the assembly were used. The resulting assembly was then further analysed by aligning all contigs with a minimal length of 1,000 nt to the *M. leprae*

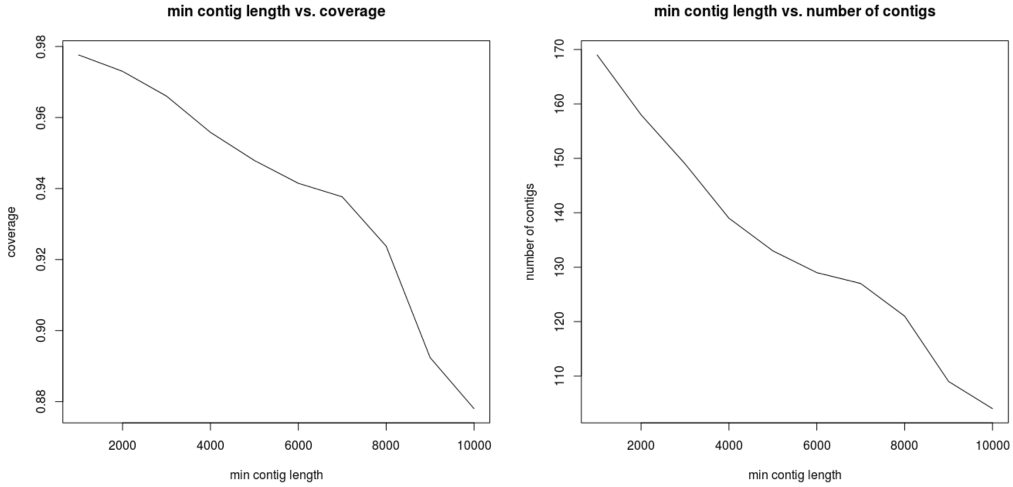


Figure 8.3.: Coverage of the *M. leprae* TN genome (left) and the number of aligned contigs (right) in relation to the minimal contig length.

TN reference genome in order to search for structural variation in the genomic architecture and to calculate the genomic coverage.

The best assembly, which was achieved with a k-mer size of 127 consisted of 18,701 contigs longer than 128 bases. However, most of the short contigs, which were not significantly longer than the k-mer size, were of low quality or contained low complexity sequences like homopolymers. The assembly was therefore filtered for contigs with a minimal length of 1,000 bases, which resulted in 2,354 metagenomic contigs. The average contig size was 6,165 bases with an N50 of 20,008 bases and a maximal contig size of 241,390 bases. Of the 2,354 metagenome contigs 169 could be aligned to the *M. leprae* TN reference genome achieving a coverage of 97.57%. Using different values for the contig length filter revealed that stricter filtering still leads to a quite good genomic coverage. Figure 8.3 shows the genomic coverage and the number of aligned contigs in relation to the minimal contig length. Up to a length threshold of 8,000 bases coverages above 90% are reached. This shows that only a very small part of the genome cannot be covered when restricting the assembly to long contigs. Furthermore, the alignment revealed that most of the gaps between contigs are due to repetitive regions. Altogether 145 of 169 contig gaps overlap a repeat region that is annotated in the genome of *M. leprae* TN. Resolving repetitive regions is a general challenge during *de novo* assembly using short reads. As most of the contig gaps are caused by repetitive regions and more than 97% of the reference genome is covered by the assembly, it can be concluded that the quality of the ancient genome assembly presented here is comparable to assemblies of modern bacterial genomes.

9. Discussion

Understanding life and the complex mechanisms by which it is sustained and by which it develops or evolves has become a scientific discipline that is more than ever based on the processing of complex information. High-throughput technologies generate huge amounts of biological data related to an organism's genome or transcriptome and thereby to the information processing of life itself. Understanding this information is therefore a crucial step towards understanding the mechanisms of life. In the past decades it has become more and more evident that even for bacteria the genomic architecture and the transcriptome are much more complex than just being a collection of genes, where each gene encodes one single protein, which fulfills one single function. Today we know that many transcripts do not encode proteins but fulfill their function as non-coding RNAs. This functionality can be regulatory or catalytic and therefore as complex as that of proteins. Even the architecture of a single transcript can be quite complex with regulatory non-coding RNA structures being part of untranslated regions of the transcript, for example. This makes clear that deciphering the genome of an organism means far more than the localization of protein-coding regions. In order to fully understand the information encrypted in a genome, more complex information has to be considered such as the localization of functional non-coding elements and the characteristics of all identified transcripts in general. This can be information about the interaction target of regulatory non-coding RNAs or about the architecture of protein-coding transcripts, e.g., the precise localization of promoter regions and transcription start sites.

In this thesis I presented several algorithms and tools for the characterization of an organism's coding and non-coding transcriptome.

With NOCORNAc a powerful and versatile software for the prediction and characterization of non-coding RNAs in bacterial genomes is provided. NOCORNAc has been applied to the identification and further assessment of non-coding RNAs in the antibiotic producing bacterium *Streptomyces coelicolor*. By the integration of high-resolution time series expression data for predicted elements a major contribution to the characterization of this bacterium's non-coding transcriptome has been made. In addition, non-coding elements potentially involved in the regulation of antibiotics production could be identified.

For the comparative analysis of genomic and transcriptomic data across several genomes the SuperGenome algorithm is introduced, which allows for the generation of a common coordinate system for multiple genomes that differ by insertions, deletions or genomic rearrangements. In this thesis the SuperGenome has been utilized as the basis for two quite different applications. First, it served as the basis for GenomeRing, a tool for the visualization of architectural differences between genomes. In 2011, this concept won the *Most Creative Algorithm Award* of the *Illumina iDEA Challenge*.

Second, the SuperGenome was integrated with an algorithm for transcription start site (TSS) prediction from RNA-seq data, which is also presented in this dissertation. The TSS prediction algorithm together with the SuperGenome allow for a comparative annotation of TSS across multiple genomes. The global TSS maps generated by this approach form the basis for a precise characterization of an organism's transcriptome including the identification of novel coding or non-coding genes.

In this thesis I describe the application of these algorithms to the comparative annotation of TSS in four *Campylobacter jejuni* strains. In this study several novel non-coding RNAs in the genomes of the four strains could be discovered including a new CRISPR locus. Furthermore, the promoter regions of many genes could be characterized in detail, which revealed variation of promoter activity between strains that depends on specific single nucleotide polymorphisms (SNPs) that are located in the promoter region.

In the emerging field of paleogenetics, which deals with the analysis of ancient DNA, the integration of transcriptomic data is not possible as RNA is not preserved in old samples. However, the reconstruction of full genomes from ancient DNA is possible and allows for comparative analyses together with modern genomes in order to elucidate an organism's evolution. In this dissertation I presented a computational pipeline for the comparative analysis of ancient and modern bacterial genomes, which is applied to the comparison of ancient and modern strains of *Mycobacterium leprae*, a bacterial pathogen causing leprosy.

Altogether, the algorithms and tools presented in this dissertation in addition to the knowledge that has been gained by their application represent a valuable contribution to the understanding of the organization of bacterial genomes and transcriptomes. In the following sections the individual contributions are discussed in detail.

9.1. nocoRNAc as a versatile toolbox for non-coding RNA prediction and characterization

In this dissertation nocoRNAc has been introduced, which is a Java program for the prediction and characterization of non-coding RNA transcripts in bacterial genomes [65]. For this, nocoRNAc incorporates methods for the prediction of transcriptional features.

For the detection of promoter regions the so-called SIDD model [15] is implemented in nocoRNAc, which is used for the localization of regions on the DNA where the separation of the duplex is favorable. These regions are called SIDD sites. The SIDD model considers the energy needed to separate the duplex but also the torsional energy for unwinding the helix and the influence of superhelical stress. The advantage of this model is its general applicability to bacterial genomes without any prior knowledge about transcription factor binding site motifs. This is especially valuable for organisms where detailed information about transcription factors are missing or where the number of transcription factors is very high like in *Streptomyces coelicolor*. In addition, many sequence motifs of binding sites are short and degenerated, which leads to high number of false positive predictions. It has been shown that SIDD sites are not only associated with transcription start sites but also other genomic features [122, 2, 176, 23]. However, their strong association with promoter regions has been proven [167, 165] and, furthermore, it could be shown that SIDD sites are overrepresented in upstream regions of regulatory genes that directly react to environmental changes [166]. Still, the occurrence of false positive predictions is likely. A notable disadvantage of SIDD sites is their size. SIDD sites can be tens of base pairs in length, which makes a precise localization of the transcription start difficult. In addition, SIDD sites are not strand-specific. For these reasons NOCORNAC does not rely on SIDD sites alone when predicting ncRNA transcripts.

For the detection of Rho-independent transcription terminator signals NOCORNAC utilizes the program TransTermHP [82]. Unlike SIDD sites the location of these terminator predictions is very precise and they are strand-specific. Thus, their combination with detected SIDD sites makes the prediction of ncRNA transcripts possible and allows for distinguishing them from *cis*-regulatory elements, as could have been shown in this dissertation. Using annotated ncRNA genes in *S. coelicolor* as a test set NOCORNAC predicts a transcript in 76% of the cases. In addition, 94% of the annotated *cis*-regulatory elements were classified correctly. However, NOCORNAC's performance on tRNAs was rather low. Only for 53% of all annotated tRNA genes a transcript was predicted. However, tRNAs are often transcribed polycistronically. Thus, they cannot be detected as individual transcripts.

In general it has to be considered that NOCORNAC can only predict transcripts, where transcription termination is induced by a Rho-independent terminator and transcripts whose transcription is terminated by the Rho protein will be missed. As it has been shown that Rho is involved in the termination of some ncRNAs [120] the integration of a feature for the prediction of Rho-dependent transcription termination should be in the focus of future development.

For the prediction of ncRNA transcripts NOCORNAC offers two strategies. One strategy is to apply the prediction to predefined loci that are provided by the user. For these loci there should preferably be some indication that they might contain a functional structured RNA. In this dissertation `RNAz` [61] was used for the genome-wide prediction of ncRNA loci. However, other approaches for the gathering of candidate loci can be used instead, such as the application of other prediction methods. A combination of multiple approaches could also be employed to increase sensitivity or specificity. In addition, it is up to the user to perform any kind of prefiltering on these regions. They can be restricted to intergenic regions, for example, in order to specifically search for *trans*-encoded ncRNAs. In general, however, one of the strengths of NOCORNAC is its ability to perform a genome-wide prediction including protein-coding regions to predict *cis*-encoded antisense RNAs.

The second strategy works without any predefined loci and predicts ncRNA candidate loci in a genome-wide manner by combining SIDD sites and terminator signals. NOCORNAC's structure conservation pipeline is applied to these candidates to determine their potential to contain structurally conserved RNA. For this, NOCORNAC uses BLAST to collect homologous sequences with an optimal evolutionary distance to the query and applies RNAz to an alignment of these sequences. The advantage of this approach is that the related organisms from which the homologous sequences are chosen do not have to defined by the user. Instead, the sequences are chosen individually for each query. Predicted ncRNA candidates are annotated with the results of the structure conservation pipeline, i.e., the RNAz P-values, which can be used for filtering.

nocoRNAc offers a variety of methods for the further characterization of predicted ncRNAs. One of them is the prediction of RNA-RNA interactions interactions by the utilization of the program IntaRNA. By this nocoRNAc is able to predict potential RNA-RNA interactions between all predicted ncRNA loci and protein-coding transcripts, which results in a genome-wide RNA-RNA interaction network. As has been shown in this dissertation the number of predictions is usually quite high also containing a lot of improbable interactions. Therefore, nocoRNAc provides several possibilities for the filtering of predicted interactions and the more detailed assessment of selected interaction candidates. The interaction profiles, which are calculated by nocoRNAc and which can be investigated in nocoRNAc's R environment, give an indication about which interaction sites are most probable considering the whole network. Furthermore, the R environment offers the functionality to perform RNA-RNA interaction predictions on specific subsets of elements with possibility to calculate z-scores and p-values to assess the significance of the predicted interactions. The results presented in section 4.4 suggest that these values are more informative than the free energy of the predicted interaction alone.

In addition to the analysis of interactions, nocoRNAc's R environment allows for an interactive and dynamic analysis of all prediction related data. Predicted elements can be filtered with respect to various properties such as the strength of SIDD sites, the confidence of their terminator signals or structure conservation potential. Furthermore, the powerful functionalities offered by R can be applied for the statistical assessment of these values or for their visualization. This makes nocoRNAc's usage much more efficient and additionally allows for a detailed customization of the output. Where the standard output of nocoRNAc is a GFF file, the R environment can be used to retrieve the sequences of all predicted elements, for example, or only those elements satisfying a specified condition. In this context the R package Bioconductor [53] has evolved to a powerful toolbox also for the analysis of sequencing data. As nocoRNAc's R environment is compatible with Bioconductor functions and data types the integrated analysis of predictions and experimental data can be performed within the environment as has been demonstrated in sections 4.1 and 4.2.

9.2. nocoRNAc uncovering putative ncRNA regulators in Streptomyces coelicolor

In this dissertation the application of nocoRNAc to the genome of *Streptomyces coelicolor* has been described. *S. coelicolor* is an important antibiotic producing model organism. To a large extent the mechanisms involved in the regulation of the production of these secondary metabolites have remained unexplored. Within the SysMO STREAM consortium *S. coelicolor* wild type and different mutant strains were grown under various conditions and the transcriptome, proteome and metabolome of the organism was profiled in unprecedented detail [109, 164, 99, 13, 153, 3]. The transcriptomics data was used to validate the expression of predicted ncRNAs and to characterize them further [65].

Of 403 predicted ncRNA transcripts that were measured 317 showed expression under the tested conditions. It turned out that the expression of a high number of predicted *cis*-encoded antisense RNAs correlates with the expression of their protein-coding target in many cases. In addition, the expression of several predicted ncRNAs in intergenic regions has been confirmed. Some of them show a clear reaction to the nutrient limitation event and might be involved in the regulation of metabolic processes.

Furthermore, a systematic study of potential RNA-RNA interactions between predicted ncRNAs and mRNAs identified an element putatively involved in the regulation of antibiotics production. For the ncRNA transcript, which was predicted upstream of the cold shock protein *csp*, an interaction with the mRNA of TetR, a global downregulator of antibiotics production, was predicted. In an earlier study by Martínes-Costa *et al.* [100] the region upstream of *csp* has been introduced into *S. coelicolor* using high copy number plasmids, which led to an upregulation of antibiotics production. The microarray screen indicated that the ncRNA is transcribed and if the predicted interaction with the TetR mRNA takes place *in vivo*, this would explain why an overrepresentation of that locus in the cell induces a stronger antibiotics production. In this case the expression of TetR would be downregulated due to the interaction and this would silence its repressive influence on antibiotics production. In this context NOCORNAC's ability to calculate z-scores and p-values for predicted interactions proved to be extremely useful, as the TetR mRNA would not be among the top-scoring candidates if only the free energy of the interaction would have been considered.

It has to be noted that the predicted interaction site between the csp-ncRNA and TetR is located in the middle of the TetR mRNA. Thus, the direct occlusion of the ribosome binding site cannot be the mechanism of regulation. However, the binding of the ncRNA might influence the secondary structure of the TetR mRNA in a way that represses translation or the degradation of the molecule is promoted. Furthermore, it has to be considered that the results presented in this thesis are purely based on *in silico* analyses, which are strengthened, however, by the study of Martínes-Costa *et al.*. As a next step the transcription of the csp-ncRNA has to be verified using primer extension, for example. The insertion of a constitutive promoter could be used then to overexpress the ncRNA, possibly confirming the promoting effect on antibiotics production. To my knowledge this would be the first known *trans*-encoded ncRNA that is increasing antibiotics production in *Streptomyces*, while one *trans*-encoded ncRNA [163] and one *cis*-asRNA [35] decreasing antibiotics production have already been described.

Non-coding RNAs in *S. coelicolor* have also been reported in earlier studies. Pánek *et al.* [114] identified 32 ncRNA of which 15 where also found in our study. In addition, Swiercz *et al.* [149] detected 9 ncRNAs of which we found 2. Later studies, however, made use of deep RNA sequencing techniques for the genome-wide experimental detection of ncRNAs in *S. coelicolor*. By this, Suess *et al.* [162] identified 63 ncRNA of which 29 are located antisense to a protein-coding gene. In a very recent study Moody *et al.* [104] identified ncRNAs in *S. coelicolor*, *S. avermitilis* and *S. venezuelae* confirming a high degree of antisense transcription in all three species. The authors could show that the expression of ncRNAs including asRNAs is highly species dependent even if the sequence is perfectly conserved in the three genomes. This leads to two possible conclusions. First, *in silico* analyses, which are based on genomic sequences, can only give an indication of where ncRNAs might be found in the target genome. They are not able to predict in which species or under which conditions the elements might be expressed. Second, RNA-seq analyses are also limited to some degree as they are only able to detect ncRNAs that are expressed in the investigated organism under the investigated conditions. Given the specific expression of many identified ncRNAs RNA-seq analyses alone are probably also not powerful enough to disclose the full ncRNA repertoire of an organism. In addition, a further characterization of experimentally identified elements is still necessary. Thus, the integration of experimental results and computational analyses promises to be fruitful, e.g., to assess the identified elements with respect to conservation of sequence and structure or with respect to their RNA-RNA interaction potential. Thus, NOCORNAc offers the integration with the results of the automated TSS prediction presented in section 6.1.2, which works on differential RNA-seq data [142] and which will be discussed in detail in section 9.4. By this, all integrated methods for ncRNA characterization can be applied to the identified elements. In addition, NOCORNAc can assist in the determination of the ncRNA transcript's 3' end by the utilization of predicted transcription terminator signals. While the differential RNA-seq technique allows for a precise localization of a transcript's 5' start, the exact localization of the 3' end remains much more challenging.

9.3. The SuperGenome concept as the basis for comparative analyses

Comparative analyses involving multiple species are increasingly important as the number of sequenced genomes is continuously rising. However, these analyses bear challenges as the compared genomes often differ with respect to their genomic architecture. Due to insertions, deletions but also genomic rearrangements, such as translocations or inversion, the genomes have different coordinate systems making the direct comparison of coordinate-based features difficult or impossible.

In chapter 5 the SuperGenome concept was presented as an approach to a solution of this problem. The SuperGenome is independent from a fixed reference genome and is computed on the basis of a multiple whole-genome alignment. It provides a common coordinate system for the aligned species and a mapping between this common coordinate system and the coordinate systems of the individual genomes. Furthermore, the SuperGenome implementation offers a variety of functions operating on the SuperGenome data structure to allow for coordinate transformations of genome annotations, such as genes or transcription start sites (TSS), or of genomic and transcriptomic data, such as RNA-seq expression graphs in single-nucleotide resolution. This can be utilized to compare the expression of homologous genes, if they have been aligned in the multiple whole-genome alignment, without the necessity of an ortholog mapping. Furthermore, it is possible to investigate conserved intergenic regions in order to discover novel coding or non-coding transcripts, for example. All these comparisons can be performed despite any architectural differences between the genomes as genomic regions that are conserved among the organisms will be assigned the same coordinates in the SuperGenome.

In principle, any software that is working on genomic data could also be applied to the SuperGenome and annotations that have been projected into its coordinate system. Standard genome browsers can be utilized, for example, to visualize genomic and transcriptomic data of different organisms as tracks in the same browser window. Elements in the different genomes that are related to each other, such as homologous genes or their TSS, share the same position in the SuperGenome and therefore they are also visualized at the same position in the genome browser, although their original genomic coordinates are completely different. E.g., an element might be located in the middle of the genome in one organism and due to a translocation the respective homologous element of another organism might be located at the end of the genome. The SuperGenome can compensate for these effects as long as the elements have been aligned in the multiple whole-genome alignment.

Two different applications of the SuperGenome approach have been presented in this thesis. In connection with GenomeRing it was applied to the comparative visualization of genomic architectures (section 5.2). The second application was its integration with an algorithm for TSS detection from RNA-seq data to allow for a cross-genome annotation and comparison of the detected elements (chapter 6). In the context of alignment visualization GenomeRing employs a different strategy in comparison to other visualization tools. The linear viewer integrated in Mauve [37], for example, visualizes conserved regions as colored blocks. For each genome these blocks are shown in the order they appear in that genome. As the order of the blocks differs between the genomes they are connected by lines and by a common color. Due to the varying position of a block between the different genomes it can therefore be quite difficult for the user to quickly identify blocks that are missing in specific genomes. Another circular viewer is Circos [85], where the block representations of the different genomes are laid out on a circle and connected by lines or ribbons. Both approaches focus on preserving the genomic architectures of the individual genomes in the visualization. GenomeRing, however, focuses on highlighting differences and similarities between the genomes by visualizing each block only once independently of the number of genomes in which the block is conserved. As a colored path representing the aligned genomes either traverse blocks or skips them, the user can immediately identify conserved blocks that can be found in all genomes or regions that are specific for only a subset of genomes. The architecture of the individual genomes is still shown as the paths connect the blocks in the order they appear in the respective genome. The application of GenomeRing to the four *Campylobacter jejuni* strains that were also subject of a comparative TSS analysis (chapter 7) demonstrated how this visualization technique can be utilized for the quick identification of architectural differences between the genomes. On the other hand GenomeRing also proved its ability to guide in-depth analyses as demonstrated by its application to the alignment of three *Helicobacter pylori* strains [142]. GenomeRing's connection to Mayday's linear genome browser and the integration of transcriptomic data allowed for multi-level inspection of the strains. Getting a global overview of architectural differences in GenomeRing the incorporation of gene annotations and results of transcriptomic analyses allows the user to quickly identify loci of interest, which can then be investigated on the level of gene clusters or single genes by GenomeRing's linkage to Mayday's genome browser. An even more detailed analysis is made possible by integrating position specific expression information in the form of RNA-seq data. This demonstrates how the SuperGenome-based visualization of genome alignments in GenomeRing complement the application of other tools, such as standard genome browsers, for a more comprehensive integrated analysis of genomic and transcriptomic data.

In the current implementation the number of genomes that can be visualized in GenomeRing is limited, which is due to the fact that in the visualization genomes are distinguished using different colors. However, this problem can be overcome in future implementations by developing aggregation techniques that allow for the grouping of genomes that are highly similar with respect to large parts of the genome. By this, differences between groups of genomes would be emphasized even more. It is likely that the SuperGenome basis of GenomeRing will prove to be very helpful for this task as similarities and differences between genomes and groups of genomes are implicitly modeled in the SuperGenome. Therefore, algorithms for the clustering and summarization of genomes and genomic regions will have to operate on the core data structure of the SuperGenome. Thus, the results of these summarization techniques will be beneficial not only for the visualization with GenomeRing but potentially in the context of all applications of the SuperGenome approach.

However, for all possible applications it has to be considered that the SuperGenome strongly depends on the alignment that is used as input. If homologous regions are not properly aligned in the input data, they will not share the same SuperGenome coordinates and comparative analyses will not be possible for those loci. Furthermore, in its current implementation the SuperGenome only allows for an injective mapping of coordinates between the SuperGenome and the original genomes. Thus, a SuperGenome position can map to only one position in any of the other genomes and duplications are therefore not modelled by this approach. Mauve [38], however, which is used for the generation of multiple whole-genome alignments is also not able to handle duplications, but there are other tools for genome alignment that can find duplications, such as `MUMmer` [87]. As gene duplication is an important evolutionary mechanism in prokaryotes and eukaryotes [84] the extension of the SuperGenome concept in this respect would be beneficial to allow for the comparative analysis of such events.

Another unsolved challenge is the 'evolution' of the SuperGenome itself. In the course of a study of a bacterial species, for example, the number of sequenced genomes of different strains might grow over time, or additional sequences have to be incorporated during follow-up studies. This brings up the question of how to extend an already computed SuperGenome. The incorporation of additional genomes in the multiple alignment might change the resulting SuperGenome coordinate system. The only exception would be the unlikely case that the additional genome has no insertions in comparison to the other genomes in the alignment. Therefore, the results of any analysis performed on the original SuperGenome would have to be transferred into the extended SuperGenome, which is currently an open problem in itself as there might be inconsistencies between the original multiple genome alignment and the extended one. A solution to this would be sequence-to-profile alignments, which are not supported by Mauve yet. If it is desired, however, to keep already assigned SuperGenome coordinates unchanged even if more sequences are added, a significantly more complex coordinate concept would be necessary that allows for the insertion of new positions between coordinates i and $i + 1$, for example. While it might be possible to employ such a complex coordinate concept within a closed software, the whole SuperGenome concept would loose its general applicability and software that is not designed to work on the basis of the SuperGenome, like standard genome browsers, could not be used anymore to inspect the results of the SuperGenome computation. Thus, the extensibility of the SuperGenome has to be a major subject during future development of the concept.

9.4. A novel SuperGenome-based TSS prediction approach applied to the comparative analysis of Campylobacter jejuni strains

Deep sequencing technologies become more and more efficient with respect to time and cost and the increasing number of RNA-seq data sets requires computational methods for high-throughput analyses, which have to be conducted in a comparative manner in many cases. One important part of RNA-seq data analysis is the detection of transcription start sites (TSS) as they allow for the more detailed characterization of annotated genes and the identification of novel coding or non-coding transcripts.

In this thesis an algorithm for the automated prediction of TSS from differential RNA-seq data (dRNA-seq) has been presented (chapter 6.1.3), which integrates the SuperGenome approach for a comparative characterization of TSS across different genomes. In previous studies genome-wide TSS maps on the basis of RNA-seq data were compiled manually or generated by semi-automated approaches, which predicted candidate loci that had to be further assessed by hand [142, 103, 4, 76, 46, 137, 173, 174]. The reproducibility of such methods is very limited in most cases and approaches including a lot of manual verification are extremely time-consuming with extensive comparative analyses being infeasible. The TSS prediction algorithm that has been developed during this thesis considers the same criteria which are used during manual TSS annotations. It thus provides detailed features for each predicted TSS, such as its expression strength, enrichment factor and classification with respect to its location relative to annotated genes. This makes the prediction procedure highly transparent allowing for an evaluation and refinement of the used parameters. A benchmark based on a global TSS map that has been manually annotated for *Helicobacter pylori* [142] resulted in a sensitivity of 82% and a precision rate of 75%. However, the performance of the method can be increased by the integration of more data, as the approach presented here allows for a comparative analysis across data sets from different genomes or cultivation conditions. Furthermore, the confidence of predictions can be increased by the incorporation of replicates.

Another fully automated method for the genome-wide prediction of TSS in single data sets has been recently presented by Schmidtke *et al.* [138, 7]. It employs a sophisticated statistical model calculating p-values for TSS candidates based on dRNA-seq data. However, such a statistical approach is less transparent as it is difficult to infer relevant properties like expression strength and enrichment factor from the computed p-values and these are therefore hard to interpret. Thus, it might be challenging for the user to understand why a prediction has been made and to decide on how to change parameters to influence the prediction. During future development a combination of both approaches would be fruitful as the TSS detection strategy presented in this thesis could be complemented with statistical confidence estimations. This would be extremely helpful especially for the evaluation of weak TSS candidates. Furthermore, the method by Schmidtke *et al.* is not designed for comparative analyses. A combination of both models would therefore also allow for an extension of the statistical assessment to the simultaneous analysis of multiple data sets using the SuperGenome approach.

For the convenient application of the TSS prediction procedure, the algorithm together with the SuperGenome computation have been integrated and complemented by a graphical user interface in the Java program TSSpredator. The program allows for an easy setup of all relevant parameters and file paths. TSSpredator can be applied to the comparative analysis of data sets from different cultivation conditions or from different organisms, which is made possible by the application of the SuperGenome approach. The results are compiled in a comprehensive MasterTable that allows for further analyses of the predicted elements. This table is complemented by several other result files, such as a table with TSS prediction statistics, normalized expression graphs and genome annotations and sequences, which have been transformed into the SuperGenome coordinate system in the case of cross-genome analyses.

TSSpredator provides various parameter presets yielding predictions with different sensitivity and specificity, since it can be difficult for the user to decide on reasonable parameters. However, an optimal decision on the parameter settings remains challenging. It could be shown in this dissertation that the normalization strategies that are applied during the preprocessing of the data can compensate for many effects that would otherwise lead to a bias in the prediction procedure. The normalization considers different numbers of sequencing reads as well as variation between the enrichment strengths. Nevertheless, an adaptation of the parameters to the specific requirements of the study is necessary in most cases. Ideally, the different parameter sets should be evaluated by the help of experimentally validated TSS. However, the evaluation of a large set of predicted TSS is very laborious and might not be feasible in the context of most studies. Thus, TSS annotations resulting from an automated genome-wide prediction procedure should still be considered as putative. Nonetheless, with the rising number of transcriptomic studies the global TSS maps produced by the approach presented in this dissertation will provide insights into similarities and differences between the transcriptome architectures of various bacteria. This will help to elucidate regulatory mechanisms involved in a multitude of biological functions, such as pathogenicity or adaptation.

In chapter 7 the application of the SuperGenome-based comparative TSS prediction procedure to four strains of the human pathogen *Campylobacter jejuni* has been presented. The analyses were based on a previously developed differential RNA-seq (dRNA-seq) approach [142], which allows for an enrichment of reads originating from the 5' end of primary transcripts.

The application of the approach resulted in the annotation of about 2000 TSS in each of the four *C. jejuni* strains. The comparative cross-genome detection of TSS revealed that a majority of the elements is conserved in multiple strains. However, many examples of TSS could be identified that showed a very specific expression in only one of the strains, although promoter sequences were often highly conserved among strains. This shows that a purely sequence-based analysis is not sufficient to assess conserved expression of genes in different organisms. In addition, it was observed that even those elements that are expressed in all strains show different levels of expression in many cases. Furthermore, on the basis of the automatically generated TSS maps several SNPs in promoter regions could be identified that could explain the loss of gene expression. In some cases these SNPs are located in the A/T-rich upstream sequence of the promoter, which suggests that this region plays an important role during transcription initiation. It has been suggested earlier that a high genetic microdiversity in *C. jejuni* allows this pathogen to flexibly adapt to changing environments [59] and the identified genes with SNPs in their promoter regions that influence their transcription might be good candidates of factors potentially involved in environment-specific adaptation.

The global TSS map also allowed for the identification and comparative analysis of non-coding RNAs in the four *C. jejuni* strains. A conservation analysis revealed that many of them are specific for *C. jejuni* or even for single strains while others are more widely conserved. One of the most interesting observations in this context was the discovery of a minimal CRISPR/Cas system, which in contrast to the systems found in other bacteria contains individual promoters for the crRNAs. Therefore, only one processing event is required during the maturation of the crRNAs. However, the roles of the identified ncRNAs and the influence of the discovered CRISPR locus are still unclear and have to be elucidated by further studies. Furthermore, the high number of detected antisense TSS suggests that *cis*-encoded antisense RNAs might be involved in various regulatory processes in *C. jejuni*.

Overall it has been shown in this thesis that global comparative TSS maps generated with the SuperGenome-based TSS detection algorithm presented here can be utilized to identify potentially regulatory elements on a genome-wide scale. In the light of the continuously increasing number of RNA-seq studies TSSpredator will provide a valuable approach to the comparative analysis among different bacteria or experimental conditions and allow for the annotation of TSS and characterization of promoter regions as well as for the identification of novel coding or non-coding transcripts.

9.5. A computational pipeline for the analysis of ancient pathogens

In comparative studies of bacterial organisms such as the one presented in chapter 7 different bacterial strains or species are compared to each other in order to explain phenotypic differences, for example, often in the context of pathogenicity. In other studies the reaction to different experimental conditions, such as nutrient limitations, is studied (see chapter 4). In the field of *paleogenetics*, however, the aim is the reconstruction of genomes from ancient material like skeletal remains in order to compare these genomic information to those of related modern organisms. This allows for gaining insights into the long-term evolution of that species and for tracking its origin.

In this dissertation the development of a computational pipeline for the analysis of ancient genomes was presented (chapter 8). It covers all relevant analysis steps from the preprocessing of the DNA-seq data, read mapping to the reference genome, genotyping and draft genome generation and alignment through to comparative SNP typing as the basis of phylogenetic analysis and, finally, SNP effect analysis. When processing ancient DNA their specific properties have to be considered. Most importantly ancient genomic fragments are usually degenerated to a great extent, which results in shorter and fewer fragments. Thus, paired-end sequencing with large insert sizes is not possible and the genomic coverage that can be achieved is significantly lower compared to DNA-seq from modern DNA.

It could be shown in this thesis that a comparative analysis of ancient samples together with modern samples can compensate for these effects to some extent. For SNPs that were clearly detected in some of the samples the thresholds for detection with respect to coverage and quality were loosened for the other samples in order to be able to detect SNPs despite some degree of uncertainty. This allowed for a significant reduction of missing data. Furthermore, the alignment-based comparative SNP calling allowed for the effective incorporation of multiple reference genomes.

The analysis pipeline has been applied to the comparison of modern and ancient *Mycobacterium leprae* strains. Altogether 755 SNPs were detected and annotated by the pipeline in the complete data set. Interestingly, it turned out that the genomic diversity between ancient and modern samples is rather low. This poses the question why leprosy was a much more devastating disease in medieval times than it is today. Considering the results of the analysis this is most certainly not due to a lower virulence of the pathogen but instead caused by other factors like host immunity or improved hygiene conditions.

In addition to the analysis of SNPs and the phylogeny of the studied samples a *de novo* assembly of one of the ancient samples was possible due to the high amount of endogenous *M. leprae* DNA. The comparison of this *de novo* assembly to a modern reference revealed that there are no structural variations between the ancient and modern strains. Furthermore, most gaps between contigs correspond to repetitive regions in the genome of *M. leprae*. Considering that bridging repetitive regions is generally a challenge in the context of short read assembly this shows that the data generated from the ancient DNA of that sample was of sufficient quality and depth for an effective reconstruction of the genome.

This study nicely shows how improved experimental protocols and bioinformatics methods allow for the study of the historical evolution of pathogens. By this, the field of paleogenetics, to which bioinformatics made and will continue to make valuable contributions, adds another dimension to comparative studies between organisms.

9.6. Conclusion

In the past years the complex structure of genomes and transcriptomes even of bacterial organisms has become more and more evident. The central dogma of molecular biology, which states that one gene encodes one single protein, which fulfills one single function has been falsified in many cases. Nowadays, the importance of non-coding RNAs, which are transcripts that do not encode proteins at all, is increasingly recognized. In addition, investigating the architecture of a transcript itself, i.e., the localization of its promoter region and transcription start site, for example, is essential in order to assess the potential function and the biological mechanism in which the respective gene is involved.

As biological data is produced in huge amounts due to new high-throughput technologies, genomic and transcriptomic analyses have to be efficiently automatised and often they have to be applied in a comparative manner integrating data for multiple genomes. In this dissertation I presented several methods for the computational characterization of non-coding but also protein-coding transcripts in bacteria. The application of these methods identified novel non-coding elements in different bacteria, some of which are potentially involved in the regulation of important mechanisms such as antibiotic production.

Thus, the algorithms and tools presented in this thesis could be utilized to gain significant insights into the organization of the genomes and transcriptomes of various bacterial organisms. In the future these methods will continue to support researchers in assessing the genomic and transcriptomic architectures of bacteria in the light of the ever growing amount and complexity of biological data.

Bibliography

[1] L. F. Abu-Qatouseh, S. V. Chinni, J. Seggewiss, R. A. Proctor, J. Brosius, T. S. Rozhdestvensky, G. Peters, C. von Eiff, and K. Becker. Identification of differentially expressed small non-protein-coding RNAs in Staphylococcus aureus displaying both the normal and the small-colony variant phenotype. *J Mol Med (Berl)*, 88(6):565–575, Jun 2010.

[2] P. Ak and C. J. Benham. Susceptibility to superhelically driven DNA duplex destabilization: a highly conserved property of yeast replication origins. *PLoS Comput Biol*, 1(1), Jun 2005.

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

[4] M. Albrecht, C. M. Sharma, R. Reinhardt, J. Vogel, and T. Rudel. Deep sequencing-based discovery of the Chlamydia trachomatis transcriptome. *Nucleic Acids Res*, 38(3):868–877, Jan 2010.

[5] A. Ali, S. C. Soares, E. Barbosa, A. R. Santos, D. Barh, S. M. Bakhtiar, S. S. Hassan, D. W. Ussery, A. Silva, A. Miyoshi, and V. Azevedo. Microbial Comparative Genomics: An Overview of Tools and Insights Into The Genus Corynebacterium. *J Bacteriol Parasitol*, 4(167):2, 2013.

[6] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215(3):403–410, Oct 1990.

[7] F. Amman, M. T. Wolfinger, R. Lorenz, I. L. Hofacker, P. F. Stadler, and S. Findeiß. TSSAR: TSS annotation regime for dRNA-seq data. *BMC Bioinformatics*, 15:89–89, 2014.

[8] S. Andrews. FastQC A Quality Control tool for High Throughput Sequence Data. `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`.

[9] S. V. Angiuoli and S. L. Salzberg. Mugsy: fast multiple alignment of closely related whole genomes. *Bioinformatics*, 27(3):334–342, Feb 2011.

[10] Athanasius F Bompfünewerer Consortium, R. Backofen, S. H. Bernhart, C. Flamm, C. Fried, G. Fritzsch, J. Hackermüller, J. Hertel, I. L. Hofacker, K. Missal, A. Mosig, S. J. Prohaska, D. Rose, P. F. Stadler, A. Tanzer, S. Washietl, and S. Will. RNAs everywhere: genome-wide annotation of structured RNAs. *J Exp Zool B Mol Dev Evol*, 308(1):1–25, Jan 2007.

[11] D. J. Bacon, R. A. Alm, L. Hu, T. E. Hickey, C. P. Ewing, R. A. Batchelor, T. J. Trust, and P. Guerry. DNA sequence and mutational analyses of the pVir plasmid of Campylobacter jejuni 81-176. *Infect Immun*, 70(11):6242–6250, Nov 2002.

[12] C. Barrandon, B. Spiluttini, and O. Bensaude. Non-coding RNAs regulating the transcriptional machinery. *Biol Cell*, 100(2):83–95, Feb 2008.

[13] F. Battke, A. Herbig, A. Wentzel, O. M. Jakobsen, M. Bonin, D. A. Hodgson, W. Wohlleben, T. E. Ellingsen, STREAM Consortium, and K. Nieselt. A technical platform for generating reproducible expression data from Streptomyces coelicolor batch cultivations. *Adv Exp Med Biol*, 696:3–15, 2011.

[14] F. Battke, S. Symons, and K. Nieselt. Mayday–integrative analytics for expression data. *BMC Bioinformatics*, 11:121–121, 2010.

[15] C. J. Benham and C. Bi. The analysis of stress-induced duplex destabilization in long genomic DNA sequences. *J Comput Biol*, 11(4):519–543, 2004.

[16] D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. GenBank. *Nucleic Acids Res*, 41(Database issue):36–42, Jan 2013.

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

[18] S. H. Bernhart, I. L. Hofacker, and P. F. Stadler. Local RNA base pairing probabilities in large sequences. *Bioinformatics*, 22(5):614–615, Mar 2006.

[19] S. H. Bernhart, I. L. Hofacker, S. Will, A. R. Gruber, and P. F. Stadler. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, 9:474, 2008.

[20] M. J. Bibb. Regulation of secondary metabolism in streptomycetes. *Curr Opin Microbiol*, 8(2):208–215, Apr 2005.

[21] J. J. Bijlsma, M. Lie-A-Ling, I. C. Nootenboom, C. M. Vandenbroucke-Grauls, and J. G. Kusters. Identification of loci essential for the growth of Helicobacter pylori under acidic conditions. *J Infect Dis*, 182(5):1566–1569, Nov 2000.

[22] M. Blanchette, W. J. Kent, C. Riemer, L. Elnitski, A. F. Smit, K. M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E. D. Green, D. Haussler, and W. Miller. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*, 14(4):708–715, Apr 2004.

[23] J. Bode, S. Winkelmann, S. Götze, S. Spiker, K. Tsutsui, C. Bi, P. A K, and C. Benham. Correlations between scaffold/matrix attachment region (S/-MAR) binding activity and DNA duplex destabilization energy. *J Mol Biol*, 358(2):597–613, Apr 2006.

[24] K. I. Bos, V. J. Schuenemann, G. B. Golding, H. A. Burbano, N. Waglechner, B. K. Coombes, J. B. McPhee, S. N. DeWitte, M. Meyer, S. Schmedes, J. Wood, D. J. Earn, D. A. Herring, P. Bauer, H. N. Poinar, and J. Krause.

A draft genome of Yersinia pestis from victims of the Black Death. *Nature*, 478(7370):506–510, Oct 2011.

[25] K. R. Bradnam, J. N. Fass, A. Alexandrov, P. Baranay, M. Bechner, I. Birol, S. Boisvert, J. A. Chapman, G. Chapuis, R. Chikhi, H. Chitsaz, W. C. Chou, J. Corbeil, C. Del Fabbro, T. R. Docking, R. Durbin, D. Earl, S. Emrich, P. Fedotov, N. A. Fonseca, G. Ganapathy, R. A. Gibbs, S. Gnerre, E. Godzaridis, S. Goldstein, M. Haimel, G. Hall, D. Haussler, J. B. Hiatt, I. Y. Ho, J. Howard, M. Hunt, S. D. Jackman, D. B. Jaffe, E. Jarvis, H. Jiang, S. Kazakov, P. J. Kersey, J. O. Kitzman, J. R. Knight, S. Koren, T. W. Lam, D. Lavenier, F. Laviolette, Y. Li, Z. Li, B. Liu, Y. Liu, R. Luo, I. Maccallum, M. D. Macmanes, N. Maillet, S. Melnikov, D. Naquin, Z. Ning, T. D. Otto, B. Paten, O. S. Paulo, A. M. Phillippy, F. Pina-Martins, M. Place, D. Przybylski, X. Qin, C. Qu, F. J. Ribeiro, S. Richards, D. S. Rokhsar, J. G. Ruby, S. Scalabrin, M. C. Schatz, D. C. Schwartz, A. Sergushichev, T. Sharpe, T. I. Shaw, J. Shendure, Y. Shi, J. T. Simpson, H. Song, F. Tsarev, F. Vezzi, R. Vicedomini, B. M. Vieira, J. Wang, K. C. Worley, S. Yin, S. M. Yiu, J. Yuan, G. Zhang, H. Zhang, S. Zhou, and I. F. Korf. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience*, 2(1):10–10, Jul 2013.

[26] S. Brantl. Regulatory mechanisms employed by cis-encoded antisense RNAs. *Curr Opin Microbiol*, 10(2):102–109, Apr 2007.

[27] A. W. Briggs, U. Stenzel, P. L. Johnson, R. E. Green, J. Kelso, K. Prüfer, M. Meyer, J. Krause, M. T. Ronan, M. Lachmann, and S. Pääbo. Patterns of damage in genomic DNA sequences from a Neandertal. *Proc Natl Acad Sci U S A*, 104(37):14616–14621, Sep 2007.

[28] A. W. Briggs, U. Stenzel, M. Meyer, J. Krause, M. Kircher, and S. Pääbo. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res*, 38(6), Apr 2010.

[29] S. W. Burge, J. Daub, R. Eberhardt, J. Tate, L. Barquist, E. P. Nawrocki, S. R. Eddy, P. P. Gardner, and A. Bateman. Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res*, 41(Database issue):226–232, Jan 2013.

[30] A. Busch, A. S. Richter, and R. Backofen. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, 24(24):2849–2856, Dec 2008.

[31] I. Caldelari, Y. Chao, P. Romby, and J. Vogel. RNA-Mediated Regulation in Pathogenic Bacteria. *Cold Spring Harb Perspect Med*, 3(9), 2013.

[32] L. Childs, Z. Nikoloski, P. May, and D. Walther. Identification and classification of ncRNA molecules using graph properties. *Nucleic Acids Res*, 37(9), May 2009.

[33] P. Cingolani, A. Platts, l. e. L. Wang, M. Coon, T. Nguyen, L. Wang, S. J. Land, X. Lu, and D. M. Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, 6(2):80–92, Apr-Jun 2012.

[34] T. L. Cover and M. J. Blaser. Helicobacter pylori in health and disease. *Gastroenterology*, 136(6):1863–1873, May 2009.

[35] D. D'Alia, K. Nieselt, S. Steigele, J. Müller, I. Verburg, and E. Takano. Noncoding RNA of glutamine synthetase I modulates antibiotic production in Streptomyces coelicolor A3(2). *J Bacteriol*, 192(4):1160–1164, Feb 2010.

[36] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, and 1000 Genomes Project Analysis Group. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156–2158, Aug 2011.

[37] A. C. Darling, B. Mau, F. R. Blattner, and N. T. Perna. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res*, 14(7):1394–1403, Jul 2004.

[38] A. E. Darling, B. Mau, and N. T. Perna. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLoS One*, 5(6), 2010.

[39] A. E. Darling, I. Miklós, and M. A. Ragan. Dynamics of genome rearrangement in bacterial populations. *PLoS Genet*, 4(7), 2008.

[40] J. I. Dasti, A. M. Tareen, R. Lugert, A. E. Zautner, and U. Gross. Campylobacter jejuni: a brief overview on pathogenicity-associated factors and disease-mediating mechanisms. *Int J Med Microbiol*, 300(4):205–211, Apr 2010.

[41] E. Deltcheva, K. Chylinski, C. M. Sharma, K. Gonzales, Y. Chao, Z. A. Pirzada, M. R. Eckert, J. Vogel, and E. Charpentier. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature*, 471(7340):602–607, Mar 2011.

[42] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*, 43(5):491–498, May 2011.

[43] M. A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloë, C. Le Gall, B. Schaëffer, S. Le Crom, M. Guedj, F. Jaffrézic, and The French StatOmique Consortium. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform*, Sep 2012.

[44] A. J. Drummond, M. A. Suchard, D. Xie, and A. Rambaut. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol Biol Evol*, 29(8):1969–1973, Aug 2012.

[45] G. Dugar, A. Herbig, K. U. Förstner, N. Heidrich, R. Reinhardt, K. Nieselt, and C. M. Sharma. High-resolution transcriptome maps reveal strain-specific regulatory features of multiple Campylobacter jejuni isolates. *PLoS Genet*, 9(5), May 2013.

[46] M. J. Filiatrault, P. V. Stodghill, C. R. Myers, P. A. Bronstein, B. G. Butcher, H. Lam, G. Grills, P. Schweitzer, W. Wang, D. J. Schneider, and S. W. Cart-

inhour. Genome-wide identification of transcriptional start sites in the plant pathogen Pseudomonas syringae pv. tomato str. DC3000. *PLoS One*, 6(12), 2011.

[47] D. E. Fouts, E. F. Mongodin, R. E. Mandrell, W. G. Miller, D. A. Rasko, J. Ravel, L. M. Brinkac, R. T. DeBoy, C. T. Parker, S. C. Daugherty, R. J. Dodson, A. S. Durkin, R. Madupu, S. A. Sullivan, J. U. Shetty, M. A. Ayodeji, A. Shvartsbeyn, M. C. Schatz, J. H. Badger, C. M. Fraser, and K. E. Nelson. Major structural differences and novel potential virulence mechanisms from the genomes of multiple campylobacter species. *PLoS Biol*, 3(1), Jan 2005.

[48] K. A. Frazer, L. Pachter, A. Poliakov, E. M. Rubin, and I. Dubchak. VISTA: computational tools for comparative genomics. *Nucleic Acids Res*, 32(Web Server issue):273–279, Jul 2004.

[49] A. Friedrich. Prediction and Systematic Comparison of Non-Coding RNAs in the Bacteria Kingdom. Master thesis, University of Tübingen, 2012.

[50] P. P. Gardner, J. Daub, J. Tate, B. L. Moore, I. H. Osuch, S. Griffiths-Jones, R. D. Finn, E. P. Nawrocki, D. L. Kolbe, S. R. Eddy, and A. Bateman. Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res*, 39(Database issue):141–145, Jan 2011.

[51] D. Gautheret and A. Lambert. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *J Mol Biol*, 313(5):1003–1011, Nov 2001.

[52] T. Geissmann, C. Chevalier, M. J. Cros, S. Boisset, P. Fechter, C. Noirot, J. Schrenzel, P. François, F. Vandenesch, C. Gaspin, and P. Romby. A search for small noncoding RNAs in Staphylococcus aureus reveals a conserved sequence motif for regulation. *Nucleic Acids Res*, 37(21):7239–7257, Nov 2009.

[53] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, 5(10), 2004.

[54] J. Georg and W. R. Hess. cis-antisense RNA, another level of gene regulation in bacteria. *Microbiol Mol Biol Rev*, 75(2):286–300, Jun 2011.

[55] M. Giangrossi, G. Prosseda, C. N. Tran, A. Brandi, B. Colonna, and M. Falconi. A novel antisense RNA regulates at transcriptional level the virulence gene icsA of Shigella flexneri. *Nucleic Acids Res*, 38(10):3362–3375, Jun 2010.

[56] A. Ginolhac, M. Rasmussen, M. T. Gilbert, E. Willerslev, and L. Orlando. mapDamage: testing for damage patterns in ancient DNA sequences. *Bioinformatics*, 27(15):2153–2155, Aug 2011.

[57] M. Gottelt, S. Kol, J. P. Gomez-Escribano, M. Bibb, and E. Takano. Deletion of a regulatory gene within the cpk gene cluster reveals novel antibacterial activity in Streptomyces coelicolor A3(2). *Microbiology*, 156(Pt 8):2343–2353, Aug 2010.

[58] R. E. Green, A. S. Malaspinas, J. Krause, A. W. Briggs, P. L. Johnson, C. Uhler, M. Meyer, J. M. Good, T. Maricic, U. Stenzel, K. Prüfer, M. Siebauer, H. A. Burbano, M. Ronan, J. M. Rothberg, M. Egholm, P. Rudan, D. Brajković, Z. Kućan, I. Gusić, M. Wikström, L. Laakkonen, J. Kelso, M. Slatkin, and S. Pääbo. A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell*, 134(3):416–426, Aug 2008.

[59] E. Gripp, D. Hlahla, X. Didelot, F. Kops, S. Maurischat, K. Tedin, T. Alter, L. Ellerbroek, K. Schreiber, D. Schomburg, T. Janssen, P. Bartholomäus, D. Hofreuter, S. Woltemate, M. Uhr, B. Brenneke, P. Grüning, G. Gerlach, L. Wieler, S. Suerbaum, and C. Josenhans. Closely related Campylobacter jejuni strains from different sources reveal a generalist rather than a specialist lifestyle. *BMC Genomics*, 12:584–584, 2011.

[60] A. R. Gruber, S. H. Bernhart, I. L. Hofacker, and S. Washietl. Strategies for measuring evolutionary conservation of RNA secondary structures. *BMC Bioinformatics*, 9:122–122, 2008.

[61] A. R. Gruber, S. Findeiß, S. Washietl, I. L. Hofacker, and P. F. Stadler. RNAz 2.0: improved noncoding RNA detection. *Pac Symp Biocomput*, pages 69–79, 2010.

[62] A. R. Gruber, R. Lorenz, S. H. Bernhart, R. Neuböck, and I. L. Hofacker. The Vienna RNA websuite. *Nucleic Acids Res*, 36(Web Server issue):70–74, Jul 2008.

[63] F. J. Grundy and T. M. Henkin. From ribosome to riboswitch: control of gene expression in bacteria by RNA structural rearrangements. *Crit Rev Biochem Mol Biol*, 41(6):329–338, Nov-Dec 2006.

[64] A. Herbig, G. Jäger, F. Battke, and K. Nieselt. GenomeRing: alignment visualization based on SuperGenome coordinates. *Bioinformatics*, 28(12):7–15, Jun 2012.

[65] A. Herbig and K. Nieselt. nocoRNAc: characterization of non-coding RNAs in prokaryotes. *BMC Bioinformatics*, 12:40–40, 2011.

[66] A. Herbig. Classification of non-coding RNAs in prokaryotes. Diplomarbeit, University of Tübingen, 2008.

[67] A. Herbig and K. Nieselt. Non-coding RNA, Prediction. In W. Dubitzky, O. Wolkenhauer, K.-H. Cho, and H. Yokota, editors, *Encyclopedia of Systems Biology*, pages 1534–1538. Springer New York, 2013.

[68] J. Hertel, I. L. Hofacker, and P. F. Stadler. SnoReport: computational identification of snoRNAs with unknown targets. *Bioinformatics*, 24(2):158–164, Jan 2008.

[69] A. Hesketh, D. Fink, B. Gust, H. U. Rexer, B. Scheel, K. Chater, W. Wohlleben, and A. Engels. The GlnD and GlnK homologues of Streptomyces coelicolor A3(2) are functionally dissimilar to their nitrogen regulatory system counterparts from enteric bacteria. *Mol Microbiol*, 46(2):319–330, Oct 2002.

[70] S. Heyne, F. Costa, D. Rose, and R. Backofen. GraphClust: alignment-free structural clustering of local RNA secondary structures. *Bioinformatics*,

28(12):224–232, Jun 2012.

[71] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie / Chemical Monthly*, 125(2):167–188, 1994.

[72] S. Hoffmann, C. Otto, S. Kurtz, C. M. Sharma, P. Khaitovich, J. Vogel, P. F. Stadler, and J. Hackermüller. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput Biol*, 5(9), Sep 2009.

[73] D. Hofreuter, V. Novik, and J. E. Galán. Metabolic diversity in Campylobacter jejuni enhances specific tissue colonization. *Cell Host Microbe*, 4(5):425–433, Nov 2008.

[74] D. Hofreuter, J. Tsai, R. O. Watson, V. Novik, B. Altman, M. Benitez, C. Clark, C. Perbost, T. Jarvie, L. Du, and J. E. Galán. Unique features of a highly pathogenic Campylobacter jejuni strain. *Infect Immun*, 74(8):4694–4707, Aug 2006.

[75] H. Ikeda, J. Ishikawa, A. Hanamoto, M. Shinose, H. Kikuchi, T. Shiba, Y. Sakaki, M. Hattori, and S. Omura. Complete genome sequence and comparative analysis of the industrial microorganism Streptomyces avermitilis. *Nat Biotechnol*, 21(5):526–531, May 2003.

[76] D. Jäger, C. M. Sharma, J. Thomsen, C. Ehlers, J. Vogel, and R. A. Schmitz. Deep sequencing analysis of the Methanosarcina mazei Gö1 transcriptome in response to nitrogen availability. *Proc Natl Acad Sci U S A*, 106(51):21878–21882, Dec 2009.

[77] G. Jäger, F. Battke, and K. Nieselt. TIALA — Time series alignment analysis. In *Biological Data Visualization (BioVis), 2011 IEEE Symposium on*, pages 55–61, 2011.

[78] S. H. Kang, J. Huang, H. N. Lee, Y. A. Hur, S. N. Cohen, and E. S. Kim. Interspecies DNA microarray analysis identifies WblA as a pleiotropic down-regulator of antibiotic biosynthesis in Streptomyces. *J Bacteriol*, 189(11):4315–4319, Jun 2007.

[79] F. V. Karginov and G. J. Hannon. The CRISPR system: small RNA-guided defense in bacteria and archaea. *Mol Cell*, 37(1):7–19, Jan 2010.

[80] M. Kawano, L. Aravind, and G. Storz. An antisense RNA controls synthesis of an SOS-induced toxin evolved from an antitoxin. *Mol Microbiol*, 64(3):738–754, May 2007.

[81] D. Kim, J. S. Hong, Y. Qiu, H. Nagarajan, J. H. Seo, B. K. Cho, S. F. Tsai, and B. Ø. Palsson. Comparative analysis of regulatory elements between Escherichia coli and Klebsiella pneumoniae by genome-wide transcription start site profiling. *PLoS Genet*, 8(8), 2012.

[82] C. L. Kingsford, K. Ayanbule, and S. L. Salzberg. Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol*, 8(2), 2007.

[83] M. Kircher. Analysis of high-throughput ancient DNA sequencing data. *Methods Mol Biol*, 840:197–228, 2012.

[84] F. A. Kondrashov. Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proc Biol Sci*, 279(1749):5048–5057, Dec 2012.

[85] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra. Circos: an information aesthetic for comparative genomics. *Genome Res*, 19(9):1639–1645, Sep 2009.

[86] C. T. Kuenne, R. Ghai, T. Chakraborty, and T. Hain. GECO–linear visualization for comparative genomics. *Bioinformatics*, 23(1):125–126, Jan 2007.

[87] S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg. Versatile and open software for comparing large genomes. *Genome Biol*, 5(2), 2004.

[88] H. N. Lee, J. Huang, J. H. Im, S. H. Kim, J. H. Noh, S. N. Cohen, and E. S. Kim. Putative TetR family transcriptional regulator SCO1712 encodes an antibiotic downregulator in Streptomyces coelicolor. *Appl Environ Microbiol*, 76(9):3039–3043, May 2010.

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

[90] R. Li, H. Zhu, J. Ruan, W. Qian, X. Fang, Z. Shi, Y. Li, S. Li, G. Shan, K. Kristiansen, S. Li, H. Yang, J. Wang, and J. Wang. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res*, 20(2):265–272, Feb 2010.

[91] U. Ligges and M. Mächler. Scatterplot3d - an R Package for Visualizing Multivariate Data. *Journal of Statistical Software*, 8(11):1–20, 2003.

[92] J. Livny, M. A. Fogel, B. M. Davis, and M. K. Waldor. sRNAPredict: an integrative computational approach to identify sRNAs in bacterial genomes. *Nucleic Acids Res*, 33(13):4096–4105, 2005.

[93] J. Livny, H. Teonadi, M. Livny, and M. K. Waldor. High-throughput, kingdom-wide prediction and annotation of bacterial non-coding RNAs. *PLoS One*, 3(9), 2008.

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

[95] T. M. Lowe and S. R. Eddy. A computational screen for methylation guide snoRNAs in yeast. *Science*, 283(5405):1168–1171, Feb 1999.

[96] R. Luo, B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, J. Tang, G. Wu, H. Zhang, Y. Shi, Y. Liu, C. Yu, B. Wang, Y. Lu, C. Han, D. W. Cheung, S. M. Yiu, S. Peng, Z. Xiaoqian, G. Liu, X. Liao, Y. Li, H. Yang, J. Wang, T. W. Lam, and J. Wang. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, 1(1):18–18, 2012.

[97] S. Maeda, M. Matsuoka, N. Nakata, M. Kai, Y. Maeda, K. Hashimoto, H. Kimura, K. Kobayashi, and Y. Kashiwabara. Multidrug resistant *Mycobacterium leprae* from patients with leprosy. *Antimicrob Agents Chemother*, 45(12):3635–3639, Dec 2001.

[98] K. S. Makarova, D. H. Haft, R. Barrangou, S. J. Brouns, E. Charpentier, P. Horvath, S. Moineau, F. J. Mojica, Y. I. Wolf, A. F. Yakunin, J. van der Oost, and E. V. Koonin. Evolution and classification of the CRISPR-Cas systems. *Nat Rev Microbiol*, 9(6):467–477, Jun 2011.

[99] J. F. Martín, F. Santos-Beneit, A. Rodríguez-García, A. Sola-Landa, M. C. Smith, T. E. Ellingsen, K. Nieselt, N. J. Burroughs, and E. M. Wellington. Transcriptomic studies of phosphate control of primary and secondary metabolism in Streptomyces coelicolor. *Appl Microbiol Biotechnol*, 95(1):61–75, Jul 2012.

[100] O. H. Martínez-Costa, M. Zalacaín, D. J. Holmes, and F. Malpartida. The promoter of a cold-shock-like gene has pleiotropic effects on Streptomyces antibiotic biosynthesis. *FEMS Microbiol Lett*, 220(2):215–221, Mar 2003.

[101] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, 20(9):1297–1303, Sep 2010.

[102] I. M. Meyer. A practical guide to the art of RNA gene prediction. *Brief Bioinform*, 8(6):396–414, Nov 2007.

[103] J. Mitschke, J. Georg, I. Scholz, C. M. Sharma, D. Dienst, J. Bantscheff, B. Voss, C. Steglich, A. Wilde, J. Vogel, and W. R. Hess. An experimentally anchored map of transcriptional start sites in the model cyanobacterium Synechocystis sp. PCC6803. *Proc Natl Acad Sci U S A*, 108(5):2124–2129, Feb 2011.

[104] M. J. Moody, R. A. Young, S. E. Jones, and M. A. Elliot. Comparative analysis of non-coding RNAs in the antibiotic-producing Streptomyces bacteria. *BMC Genomics*, 14(1):558–558, Aug 2013.

[105] U. Mückstein, H. Tafer, J. Hackermüller, S. H. Bernhart, P. F. Stadler, and I. L. Hofacker. Thermodynamics of RNA-RNA binding. *Bioinformatics*, 22(10):1177–1182, May 2006.

[106] A. Muffler, D. Fischer, and R. Hengge-Aronis. The RNA-binding protein HF-I, known as a host factor for phage Qbeta RNA replication, is essential for rpoS translation in Escherichia coli. *Genes Dev*, 10(9):1143–1151, May 1996.

[107] J. W. Nicol, G. A. Helt, S. G. Blanchard, A. Raja, and A. E. Loraine. The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. *Bioinformatics*, 25(20):2730–2731, Oct 2009.

[108] C. B. Nielsen, M. Cantor, I. Dubchak, D. Gordon, and T. Wang. Visualizing genomes: techniques and challenges. *Nat Methods*, 7(3 Suppl):5–5, Mar 2010.

[109] K. Nieselt, F. Battke, A. Herbig, P. Bruheim, A. Wentzel, Ø. M. Jakobsen, H. Sletta, M. T. Alam, M. E. Merlo, J. Moore, W. A. Omara, E. R. Morrissey, M. A. Juarez-Hermosillo, A. Rodríguez-García, M. Nentwich, L. Thomas, M. Iqbal, R. Legaie, W. H. Gaze, G. L. Challis, R. C. Jansen, L. Dijkhuizen, D. A. Rand, D. L. Wild, M. Bonin, J. Reuther, W. Wohlleben, M. C. Smith, N. J. Burroughs, J. F. Martín, D. A. Hodgson, E. Takano, R. Breitling, T. E. Ellingsen, and E. M. Wellington. The dynamic architecture of the metabolic

switch in Streptomyces coelicolor. *BMC Genomics*, 11:10–10, 2010.

[110] K. Nieselt and A. Herbig. Non-coding RNA, Classification. In W. Dubitzky, O. Wolkenhauer, K.-H. Cho, and H. Yokota, editors, *Encyclopedia of Systems Biology*, pages 1532–1534. Springer New York, 2013.

[111] E. Nudler and A. S. Mironov. The riboswitch control of bacterial metabolism. *Trends Biochem Sci*, 29(1):11–17, Jan 2004.

[112] Y. Ohnishi, J. Ishikawa, H. Hara, H. Suzuki, M. Ikenoya, H. Ikeda, A. Yamashita, M. Hattori, and S. Horinouchi. Genome sequence of the streptomycin-producing microorganism Streptomyces griseus IFO 13350. *J Bacteriol*, 190(11):4050–4060, Jun 2008.

[113] H. Pages, P. Aboyoun, R. Gentleman, and S. DebRoy. *Biostrings: String objects representing biological sequences, and matching algorithms.* R package version 2.22.0.

[114] J. Pánek, J. Bobek, K. Mikulík, M. Basler, and J. Vohradský. Biocomputational prediction of small non-coding RNAs in Streptomyces. *BMC Genomics*, 9:217–217, 2008.

[115] A. Paradkar, A. Trefzer, R. Chakraburtty, and D. Stassi. Streptomyces genetics: a genomic perspective. *Crit Rev Biotechnol*, 23(1):1–27, 2003.

[116] C. T. Parker, B. Quiñones, W. G. Miller, S. T. Horn, and R. E. Mandrell. Comparative genomic analysis of Campylobacter jejuni strains reveals diversity due to genomic elements similar to those present in C. jejuni strain RM1221. *J Clin Microbiol*, 44(11):4125–4135, Nov 2006.

[117] K. Pawlik, M. Kotowska, K. F. Chater, K. Kuczek, and E. Takano. A cryptic type I polyketide synthase (cpk) gene cluster in Streptomyces coelicolor A3(2). *Arch Microbiol*, 187(2):87–99, Feb 2007.

[118] K. Pawlik, M. Kotowska, and P. Kolesiński. Streptomyces coelicolor A3(2) produces a new yellow pigment associated with the polyketide synthase Cpk. *J Mol Microbiol Biotechnol*, 19(3):147–151, 2010.

[119] J. S. Pedersen, G. Bejerano, A. Siepel, K. Rosenbloom, K. Lindblad-Toh, E. S. Lander, J. Kent, W. Miller, and D. Haussler. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS computational biology*, 2(4):e33, 2006.

[120] J. M. Peters, R. A. Mooney, P. F. Kuan, J. L. Rowland, S. Keles, and R. Landick. Rho directs widespread termination of intragenic and stable RNA transcription. *Proc Natl Acad Sci U S A*, 106(36):15406–15411, Sep 2009.

[121] C. Pichon and B. Felden. Small RNA genes expressed from Staphylococcus aureus genomic and pathogenicity islands with specific expression among pathogenic strains. *Proc Natl Acad Sci U S A*, 102(40):14249–14254, Oct 2005.

[122] Z. Polonskaya, C. J. Benham, and J. Hearing. Role for a region of helically unstable DNA within the Epstein-Barr virus latent cycle origin of DNA replication oriP in origin function. *Virology*, 328(2):282–291, Oct 2004.

[123] O. Popa and T. Dagan. Trends and barriers to lateral gene transfer in prokaryotes. *Curr Opin Microbiol*, 14(5):615–623, Oct 2011.

[124] K. Prüfer, U. Stenzel, M. Hofreiter, S. Pääbo, J. Kelso, and R. E. Green. Computational challenges in the analysis of ancient DNA. *Genome Biol*, 11(5), 2010.

[125] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[126] M. Rehmsmeier, P. Steffen, M. Hochsmann, and R. Giegerich. Fast and effective prediction of microRNA/target duplexes. *RNA*, 10(10):1507–1517, Oct 2004.

[127] J. P. Richardson. Loading Rho to terminate transcription. *Cell*, 114(2):157–159, Jul 2003.

[128] S. Rigali, H. Nothaft, E. E. Noens, M. Schlicht, S. Colson, M. Müller, B. Joris, H. K. Koerten, D. A. Hopwood, F. Titgemeyer, and G. P. van Wezel. The sugar phosphotransferase system of Streptomyces coelicolor is regulated by the GntR-family regulator DasR and links N-acetylglucosamine metabolism to the control of development. *Mol Microbiol*, 61(5):1237–1251, Sep 2006.

[129] E. Rivas, R. J. Klein, T. A. Jones, and S. R. Eddy. Computational identification of noncoding RNAs in E. coli by comparative genomics. *Current Biology*, 11(17):1369–1373, 2001.

[130] M. D. Robinson and A. Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*, 11(3), 2010.

[131] A. Rodríguez-García, C. Barreiro, F. Santos-Beneit, A. Sola-Landa, and J. F. Martín. Genome-wide transcriptomic and proteomic analysis of the primary response to phosphate limitation in Streptomyces coelicolor M145 and in a Δ*phoP* mutant. *Proteomics*, 7(14):2410–2429, Jul 2007.

[132] P. Saetrom, R. Sneve, K. I. Kristiansen, O. Snove, T. Grünfeld, T. Rognes, and E. Seeberg. Predicting non-coding RNA genes in Escherichia coli with boosted genetic programming. *Nucleic Acids Res*, 33(10):3263–3270, 2005.

[133] R. Salari, C. Aksay, E. Karakoc, P. J. Unrau, I. Hajirasouliha, and S. C. Sahinalp. smyRNA: a novel Ab initio ncRNA gene finder. *PLoS One*, 4(5), 2009.

[134] S. L. Salzberg, A. M. Phillippy, A. Zimin, D. Puiu, T. Magoc, S. Koren, T. J. Treangen, M. C. Schatz, A. L. Delcher, M. Roberts, G. Marçais, M. Pop, and J. A. Yorke. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res*, 22(3):557–567, Mar 2012.

[135] P. Schattner, A. N. Brooks, and T. M. Lowe. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res*, 33(Web Server issue):686–689, Jul 2005.

[136] P. Schattner, W. A. Decatur, C. A. Davis, M. Ares, M. J. Fournier, and T. M. Lowe. Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the Saccharomyces cerevisiae genome. *Nucleic Acids Res*, 32(14):4281–4296, 2004.

[137] J. P. Schlüter, J. Reinkensmeier, S. Daschkey, E. Evguenieva-Hackenberg, S. Janssen, S. Jänicke, J. D. Becker, R. Giegerich, and A. Becker. A genome-

wide survey of sRNAs in the symbiotic nitrogen-fixing alpha-proteobacterium Sinorhizobium meliloti. *BMC Genomics*, 11:245–245, 2010.

[138] C. Schmidtke, S. Findeiss, C. M. Sharma, J. Kuhfuss, S. Hoffmann, J. Vogel, P. F. Stadler, and U. Bonas. Genome-wide transcriptome analysis of the plant pathogen Xanthomonas identifies sRNAs with putative virulence functions. *Nucleic Acids Res*, 40(5):2020–2031, Mar 2012.

[139] M. Schubert, A. Ginolhac, S. Lindgreen, J. F. Thompson, K. A. Al-Rasheid, E. Willerslev, A. Krogh, and L. Orlando. Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics*, 13:178–178, 2012.

[140] V. J. Schuenemann, K. Bos, S. DeWitte, S. Schmedes, J. Jamieson, A. Mittnik, S. Forrest, B. K. Coombes, J. W. Wood, D. J. Earn, W. White, J. Krause, and H. N. Poinar. Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of Yersinia pestis from victims of the Black Death. *Proc Natl Acad Sci U S A*, 108(38):746–752, Sep 2011.

[141] V. J. Schuenemann, P. Singh, T. A. Mendum, B. Krause-Kyora, G. Jäger, K. I. Bos, A. Herbig, C. Economou, A. Benjak, P. Busso, A. Nebel, J. L. Boldsen, A. Kjellström, H. Wu, G. R. Stewart, G. M. Taylor, P. Bauer, O. Y. Lee, H. H. Wu, D. E. Minnikin, G. S. Besra, K. Tucker, S. Roffey, S. O. Sow, S. T. Cole, K. Nieselt, and J. Krause. Genome-wide comparison of medieval and modern Mycobacterium leprae. *Science*, 341(6142):179–183, Jul 2013.

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

[143] J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. Jones, and I. Birol. ABySS: a parallel assembler for short read sequence data. *Genome Res*, 19(6):1117–1123, Jun 2009.

[144] D. D. Sledjeski, C. Whitman, and A. Zhang. Hfq is necessary for regulation by the untranslated RNA DsrA. *J Bacteriol*, 183(6):1997–2005, Mar 2001.

[145] W. J. Snelling, M. Matsuda, J. E. Moore, and J. S. Dooley. Campylobacter jejuni. *Lett Appl Microbiol*, 41(4):297–302, 2005.

[146] G. Soldà, I. V. Makunin, O. U. Sezerman, A. Corradin, G. Corti, and A. Guffanti. An Ariadne's thread to the identification and annotation of noncoding RNAs in eukaryotes. *Brief Bioinform*, 10(5):475–489, Sep 2009.

[147] J. Sridhar, N. Sambaturu, S. R. Narmada, R. Sabarinathan, H. Y. Ou, Z. Deng, K. Sekar, Z. A. Rafi, and K. Rajakumar. sRNAscanner: a computational tool for intergenic small RNA detection in bacterial genomes. *PLoS One*, 5(8), 2010.

[148] S. Sun, R. Ke, D. Hughes, M. Nilsson, and D. I. Andersson. Genome-wide detection of spontaneous chromosomal rearrangements in bacteria. *PLoS One*, 7(8), 2012.

[149] J. P. Swiercz, Hindra, J. Bobek, H. J. Haiser, C. Di Berardo, B. Tjaden, and M. A. Elliot. Small non-coding RNAs in Streptomyces coelicolor. *Nucleic Acids Res*, 36(22):7240–7251, Dec 2008.

[150] S. Symons, C. Zipplies, F. Battke, and K. Nieselt. Integrative systems biology visualization with MAYDAY. *J Integr Bioinform*, 7(3), 2010.

[151] K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, and S. Kumar. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol*, 28(10):2731–2739, Oct 2011.

[152] M. P. Terns and R. M. Terns. CRISPR-based adaptive immune systems. *Curr Opin Microbiol*, 14(3):321–327, Jun 2011.

[153] L. Thomas, D. A. Hodgson, A. Wentzel, K. Nieselt, T. E. Ellingsen, J. Moore, E. R. Morrissey, R. Legaie, STREAM Consortium, W. Wohlleben, A. Rodríguez-García, J. F. Martín, N. J. Burroughs, E. M. Wellington, and M. C. Smith. Metabolic switches and adaptations deduced from the proteomes of Streptomyces coelicolor wild type and phoP mutant grown in batch culture. *Mol Cell Proteomics*, 11(2), Feb 2012.

[154] J. D. Thompson, T. J. Gibson, and D. G. Higgins. Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics*, Chapter 2:Unit 2.3, Aug 2002.

[155] B. Tjaden. Prediction of small, noncoding RNAs in bacteria using heterogeneous data. *J Math Biol*, 56(1-2):183–200, Jan 2008.

[156] T. T. Tran, F. Zhou, S. Marshburn, M. Stead, S. R. Kushner, and Y. Xu. De novo computational prediction of non-coding RNA genes in prokaryotic genomes. *Bioinformatics*, 25(22):2897–2905, Nov 2009.

[157] H. H. Tseng, Z. Weinberg, J. Gore, R. R. Breaker, and W. L. Ruzzo. Finding non-coding RNAs through genome-scale clustering. *J Bioinform Comput Biol*, 7(2):373–388, Apr 2009.

[158] I. Uchiyama, T. Higuchi, and I. Kobayashi. CGAT: a comparative genome analysis tool for visualizing alignments in the analysis of complex evolutionary changes between closely related genomes. *BMC Bioinformatics*, 7:472–472, 2006.

[159] C. Unoson and E. G. Wagner. A small SOS-induced toxin is targeted against the inner membrane in Escherichia coli. *Mol Microbiol*, 70(1):258–270, Oct 2008.

[160] A. V. Uzilov, J. M. Keegan, and D. H. Mathews. Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. *BMC Bioinformatics*, 7:173–173, 2006.

[161] L. Van Melderen. Toxin-antitoxin systems: why so many, what for? *Curr Opin Microbiol*, 13(6):781–785, Dec 2010.

[162] M. P. Vockenhuber, C. M. Sharma, M. G. Statt, D. Schmidt, Z. Xu, S. Dietrich, H. Liesegang, D. H. Mathews, and B. Suess. Deep sequencing-based identification of small non-coding RNAs in Streptomyces coelicolor. *RNA Biol*, 8(3):468–477, May-Jun 2011.

[163] M. P. Vockenhuber and B. Suess. Streptomyces coelicolor sRNA scr5239 inhibits agarase expression by direct base pairing to the dagA coding region. *Microbiology*, 158(Pt 2):424–435, Feb 2012.

[164] E. Waldvogel, A. Herbig, F. Battke, R. Amin, M. Nentwich, K. Nieselt, T. E. Ellingsen, A. Wentzel, D. A. Hodgson, W. Wohlleben, and Y. Mast. The PII protein GlnK is a pleiotropic regulator for morphological differentiation and secondary metabolism in Streptomyces coelicolor. *Appl Microbiol Biotechnol*, 92(6):1219–1236, Dec 2011.

[165] H. Wang and C. J. Benham. Promoter prediction and annotation of microbial genomes based on DNA sequence and structural responses to superhelical stress. *BMC Bioinformatics*, 7:248–248, 2006.

[166] H. Wang and C. J. Benham. Superhelical destabilization in regulatory regions of stress response genes. *PLoS Comput Biol*, 4(1), Jan 2008.

[167] H. Wang, M. Noordewier, and C. J. Benham. Stress-induced DNA duplex destabilization (SIDD) in the E. coli genome: SIDD sites are closely associated with promoters. *Genome Res*, 14(8):1575–1584, Aug 2004.

[168] T. M. Wassenaar and M. J. Blaser. Pathophysiology of Campylobacter jejuni infections of humans. *Microbes Infect*, 1(12):1023–1033, Oct 1999.

[169] L. S. Waters and G. Storz. Regulatory RNAs in bacteria. *Cell*, 136(4):615–628, Feb 2009.

[170] S. Will, T. Joshi, I. L. Hofacker, P. F. Stadler, and R. Backofen. LocARNA-P: Accurate boundary prediction and improved detection of structural RNAs. *RNA*, 18(5):900–914, 2012.

[171] E. Willerslev, A. J. Hansen, R. Ronn, T. B. Brand, I. Barnes, C. Wiuf, D. Gilichinsky, D. Mitchell, and A. Cooper. Long-term persistence of bacterial DNA. *Curr Biol*, 14(1):9–10, Jan 2004.

[172] K. S. Wilson and P. H. von Hippel. Transcription termination at intrinsic terminators: the role of the RNA hairpin. *Proc Natl Acad Sci U S A*, 92(19):8793–8797, Sep 1995.

[173] O. Wurtzel, R. Sapra, F. Chen, Y. Zhu, B. A. Simmons, and R. Sorek. A single-base resolution map of an archaeal transcriptome. *Genome Res*, 20(1):133–141, Jan 2010.

[174] O. Wurtzel, N. Sesto, J. R. Mellin, I. Karunker, S. Edelheit, C. Bécavin, C. Archambaud, P. Cossart, and R. Sorek. Comparative transcriptomics of pathogenic and non-pathogenic Listeria species. *Mol Syst Biol*, 8:583–583, 2012.

[175] N. Yachie, K. Numata, R. Saito, A. Kanai, and M. Tomita. Prediction of non-coding and antisense RNA genes in Escherichia coli with Gapped Markov Model. *Gene*, 372:171–181, May 2006.

[176] M. P. Yadav, S. Padmanabhan, V. P. Tripathi, R. K. Mishra, and D. D. Dubey. Analysis of stress-induced duplex destabilization (SIDD) properties of replication origins, genes and intergenes in the fission yeast, Schizosaccharomyces pombe. *BMC Res Notes*, 5:643–643, 2012.

[177] J. H. Yang, X. C. Zhang, Z. P. Huang, H. Zhou, M. B. Huang, S. Zhang, Y. Q. Chen, and L. H. Qu. snoSeeker: an advanced computational package for screening of guide and orphan snoRNA genes in the human genome. *Nucleic Acids Res*, 34(18):5112–5123, 2006.

[178] Z. Yao, Z. Weinberg, and W. L. Ruzzo. CMfinder—a covariance model based RNA motif finding algorithm. *Bioinformatics*, 22(4):445–452, 2006.

[179] K. T. Young, L. M. Davis, and V. J. Dirita. Campylobacter jejuni: molecular biology and pathogenesis. *Nat Rev Microbiol*, 5(9):665–679, Sep 2007.

[180] yWorks GmbH. The yEd Graph Editor. `http://www.yworks.com/en/products_yed_about.html`.

[181] D. R. Zerbino and E. Birney. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res*, 18(5):821–829, May 2008.

[182] A. Zhang, S. Altuvia, A. Tiwari, L. Argaman, R. Hengge-Aronis, and G. Storz. The OxyS regulatory RNA represses rpoS translation and binds the Hfq (HF-I) protein. *EMBO J*, 17(20):6061–6068, Oct 1998.

A. Publications

A.1. Articles

2010

- **Kay Nieselt, Florian Battke, Alexander Herbig**, Per Bruheim, Alexander Wentzel, Øyvind Jakobsen, Håvard Sletta, Tauqueer Alam, Elena Merlo, Jay Moore, Walid Omara, Edward Morrissey, Miguel Juarez-Hermosillo, Lubbert Dijkhuizen, David Rand, David Wild, Michael Bonin, Jens Reuther, Wolfgang Wohlleben, Margaret Smith, Nigel Burroughs, Juan Martín, David Hodgson, Eriko Takano, Rainer Breitling, Trond Ellingsen, and Elizabeth Wellington. *The dynamic architecture of the metabolic switch in Streptomyces coelicolor.* BMC Genomics 2010, 11:10

2011

- **Florian Battke, Alexander Herbig**, Alexander Wentzel, Øyvind M. Jakobsen, Michael Bonin, Dave A. Hodgson, Wolfgang Wohlleben, Trond E. Ellingsen, and Kay Nieselt. *A Technical Platform for Generating Reproducible Expression Data from Streptomyces coelicolor Batch Cultivations.* Advances in Experimental Medicine and Biology: Software Tools and Algorithms for Biological Systems
- **Alexander Herbig**, and Kay Nieselt. *nocoRNAc: characterization of non-coding RNAs in prokaryotes.* BMC Bioinformatics 2011, 12:40
- **Florian Battke, Stephan Symons**, Alexander Herbig, and Kay Nieselt. *GaggleBridge: Collaborative data analysis.* Bioinformatics 2011, 27 (18): 2612–2613
- **Eva Waldvogel, Alexander Herbig**, Florian Battke, Merle Nentwich, Kay Nieselt, Trond E. Ellingsen, David A. Hodgson, Wolfgang Wohlleben, and Yvonne Mast. *The P_{II} protein GlnK is a pleiotropic regulator for morphological differentiation and secondary metabolism in Streptomyces coelicolor.* Applied Microbiology and Biotechnology 2011: 1–18

2012

- **Alexander Herbig, Florian Battke, Günter Jäger**, and Kay Nieselt. *GenomeRing: alignment visualization in SuperGenome coordinates.* Proceedings of the ISMB, 2012

2013

- **Kay Nieselt**, and Alexander Herbig. *Non-coding RNA Classification.* Encyclopedia of Systems Biology 2013, 1532-1534
- **Alexander Herbig**, and Kay Nieselt. *Non-coding RNA Prediction.* Encyclopedia of Systems Biology 2013, 1534-1538
- **Christopher F. Schuster**, Jung-Ho Park, Marcel Prax, Alexander Herbig, Kay Nieselt, Ralf Rosenstein, Masayori Inouye, and Ralph Bertram. *Characterization of a mazEF toxin-antitoxin homologue from Staphylococcus equorum.* J Bacteriol. 2013, Jan; 195(1): 115-25
- **Gaurav Dugar, Alexander Herbig**, Konrad U. Förstner, Nadja Heidrich, Richard Reinhardt, Kay Nieselt, and Cynthia M. Sharma. *High-resolution transcriptome maps reveal strain-specific regulatory features of multiple Campylobacter jejuni isolates.* PLoS Genet. 2013, May; 9(5): e1003495
- **Verena J. Schuenemann, Pushpendra Singh, Thomas A. Mendum, Ben Krause-Kyora, Günter Jäger, Kirsten I. Bos**, Alexander Herbig, Christos Economou, Andrej Benjak, Philippe Busso, Almut Nebel, Jesper L. Boldsen, Anna Kjellström, Huihai Wu, Graham R. Stewart, G. Michael Taylor, Peter Bauer, Oona Y.-C. Lee, Houdini H.T. Wu, David E. Minnikin, Gurdyal S. Besra, Katie Tucker, Simon Roffey, Samba O. Sow, Stewart T. Cole, Kay Nieselt, and Johannes Krause. *Genome-wide comparison of medieval and modern Mycobacterium leprae.* Science 2013, Jul 12; 341(6142): 179-83

A.2. Posters & Presentations

2009

- Alexander Herbig, and Kay Nieselt. *Characterisation of non-coding RNAs and RNA-RNA interactions in S. coelicolor.* **Poster** at the German Conference on Bioinformatics 2009

2010

- Alexander Herbig, Florian Battke, and Kay Nieselt. *Integrative Transcriptome Analysis of Streptomyces coelicolor.* **Presentation** at the International VAAM-Workshop 2010
- Alexander Herbig, and Kay Nieselt. *Characterization of Non-Coding RNAs in Streptomyces coelicolor.* **Poster** at the EMBO—EMBL Symposium: The Non-Coding Genome 2010

2011

- Alexander Herbig, Cynthia Sharma, and Kay Nieselt. *Prediction and Annotation of Transcription Start Sites and Non-Coding RNAs from RNA-Seq Data.* **Poster** at the CeBiTec Symposium 2011
- Alexander Herbig, and Kay Nieselt. *Characterization of the Non-Coding RNA Transcriptome of Streptomyces coelicolor under different Nutrient Limitations.* **Poster** at the Prokagenomics 2011

2012

- Alexander Herbig, Florian Battke, Günter Jäger, and Kay Nieselt. *GenomeRing: alignment visualization in SuperGenome coordinates.* **Presentation** at the ISMB 2012