Unveiling the global patterns of diurnally oscillating genes in a wild tobacco *Nicotiana attenuata*

by

Wencke Walter

Thesis submitted to the Faculty of mathematics and computer science
of Friedrich-Schiller University of Jena
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE
in
Bioinformatics

Prof. Dr. Stefan Schuster
Prof. Dr. Ian T. Baldwin

Jena, 05.09.2013

Table of contents

Abstract

This study was performed with three goals in mind: (1) to develop an efficient method to detect oscillating genes in the genome of *Nicotiana attenuata*, (2) to estimate the influence of light on the expression of these genes, and (3) to test the assumption that the circadian clock is organ specific.

Three different datasets were used to analyze the gene expression in two organisms, tissues and conditions. The datasets contain time-series with 6 to 13 time-points sampled in a 4 hour interval over two days. Based on the original data new time-series were generated by randomly selecting one of the three biological replicates at each time point. Two pre-developed algorithms (ARSER and HAYSTACK) were used to detect rhythmically expressed genes. These genes were classified into different ZT/CT groups according to their molecular peaking time. By comparing ZT/CT groups of a given gene under different conditions or in various tissues, changes in gene expression could be determined.

First of all, the results of this research show that a randomization step with replicated microarray data is essential to detect oscillating genes with a low number of false positives. The results suggest that on average 13.4% of the analyzed genes seem to be regulated by the circadian clock. Furthermore, the molecular peaking time changes depending on the prevailing conditions. When plants transferred to different photocycles, they synchronize their gene expression with the environment in anticipation of dawn and dusk. On the other hand, each cell contains an individual endogenous oscillator and the assumption that the plant clock is organ-specific could be confirmed. Moreover, it has been found, that the rhythm of oscillating genes persists under constant light conditions, but the period is shortened and with increasing time the waveform broadens until finally the rhythm gradually damps out.

Statement of Affirmation

I hereby declare that the master thesis submitted was in all parts exclusively prepared on my own, and that other resources or other means (including electronic media and online sources), than those explicitly referred to, have not been utilized. All implemented fragments of text, employed in a literal and/or analogous manner, have been marked as such.

Jena, September 2013

Acknowledgment

First of all I would like to express my sincere gratitude to Prof. Dr. Stefan Schuster and Prof. Dr. Ian T. Baldwin for allowing me to conduct this research under their auspices. I am especially grateful for their confidence and the freedom they gave me to do this work.

I would like to show my greatest appreciation to Assoc. Prof. Dr. Ines Heiland and Dr. Sang-Gyu Kim. I can't say thank you enough for the useful comments, remarks and engagement through the learning process of this master thesis. I felt motivated and encouraged every time I attended the meetings or Skype sessions.

I extend my sincere thanks to all members of the Circadian Clock group, and all those who contributed directly or indirectly to this thesis. Thanks for your expertise and helpful suggestions whenever I had problems during the study.

I dedicate this thesis to my parents who unremittingly supported me during my years of study. They made this work possible.

List of Figures

Chapter 5

List of abbreviations

The following table describes the significance of various abbreviations and acronyms used throughout the thesis. The page on which each one is defined or first used is also given.

Abbreviation	Meaning	Page
AR	Autoregressive model	26
ARSER_20	List of genes detected by ARSER as rhythmically expressed in at least 20 out of 30 time-series	43
CT	Circadian time	18
Gr	Group	51
HAYSTACK_20	List of genes detected by HAYSTACK as rhythmically expressed in at least 20 out of 30 time-series	43
h	hour	21
LD	Light/dark cycle	16
LD_20	List of genes detected by ARSER as rhythmically expressed in at least 20 out of 30 time-series under light/dark cycles	43
Leaf_20	List of genes detected by ARSER as rhythmically expressed in at least 20 out of 30 time-series under light/dark cycles in the leaf tissue	61
LL	Constant light conditions	18
LL_20	List of genes detected by ARSER as rhythmically expressed in at least 20 out of 30 time-series under constant light condition	43
Log2	Binary logarithm	25
MMC-β	Multiple-measures corrected β values	20
Root_20	List of genes detected by ARSER as rhythmically expressed in at least 20 out of 30 time-series under light/dark cycles in the root tissue	61
ZT	Zeitgeber time	18
12:12	12 hour light and 12 hour dark	55
16:8	16 hour light and 8 hour dark	55

List of tables

1. Introduction

In the year 1492 Christopher Columbus discovered the "New World" and during his second landfall at Bariay, Cuba he sent men to investigate the country. Their mission was to find the Emperor of China, but instead of a palace they found a native Taino village and were the first to observe the smoking of tobacco. It didn't take long that the whole crew got used to this new habit. Beside exotic comestibles and objects Columbus also brought the tobacco plant to Europe and following the natural behavior of skepticism towards new things it took some time until the public accepted it. Jean Nicot, a French doctor and an ambassador for his country in Lisbon, attributed a healing character to the plant, which give rise to the culture of snuffing and the plant's future popularity throughout Europe. He could observe, that treating wounds with tobacco leaves promotes the healing and also headaches could be eased. Excited by the effect of the plant and by knowing that the Queen of France, Catherine de Medici, suffers from chronic migraine headaches Nicot sent her some tobacco seeds. Catherine de Medici became convinced that tobacco had healing properties and by receiving instructions from Nicot she got used to the habit of snuffing. Because of its preventative role of the plant the Queen and Nicot made it publicly respectable in France. From now on members of the French court used the tobacco powder to stave off various illnesses and it's likely that many users developed addictions to it. In recognition of Nicot's role in popularizing the plant, the French botanist Jacques Dalechamps named it *herba nicotiana* in the year 1586. In 1828 German chemists from Heidelberg were able to isolate the active alkaloid in the tobacco plant leaves for the first time and named it *nicotine*. The genus *Nicotiana* contains many different species which differ in morphology and dispersal. One of these species domiciled in the southwest of the USA was named *Nicotiana*

Figure 1.1: Stamp issued in 1961 the 400th anniversary of the introduction of tobacco into France [25].

attenuata by John Torrey because of their long and elongated leaves and flowers. The Latin word “attenuata” means “thin or weak”. It’s an annual herb and occupies a special ecological niche. *N. attenuata* prefers a post-fire environment with its nitrogen-rich soil. A dormant seed bank is established during post-fire period. The plants synchronize their germination from the seed bank with the post-fire environment. Stimulants in wood smoke give the signal for the germination of the seeds. Forest fires do not follow any known rule and as a result the appearance of *N. attenuata* is hard to predict.

1.1 Motivation

At the first glance the annual herb *Nicotiana attenuata* may not appear extraordinary. It is a common plant exceeding a meter in maximum height that is easily recognized by its glandular foliage and white, tubular flowers. The plant does not impose by its ordinary appearance, that’s for sure, but its elaborated responses to predation and its exceptional habitat justify a detailed investigation.

It is assumed that a remarkable part of plant defense is regulated by the circadian clock. For *Arabidopsis thaliana* the molecular basis of circadian rhythms is known but although oscillator mechanisms are conserved through evolution, actual clock components do not seem to be [53]. To figure out the single components of a network it is necessary to research on a large number of genes simultaneously. Microarray technology allows to monitor the expression of hundreds or thousands of genes in a single reaction quickly and in an efficient manner. Thus with the help of microarrays we should be able to gain some new information about the global patterns of diurnally oscillating genes in *Nicotiana attenuata*.

1.2 Aims

The genome of *Nicotiana attenuata* was recently annotated and the aim of this work is to precisely define the genomic landscape. The annotation of a genome is not the satisfactory ending of a long row of experiments but the beginning of a much bigger scientific challenge.

"Denn ein Buch zu lesen, das in einer Sprache geschrieben ist, deren Alphabet man zwar kennt, aber dessen Worte und deren Bedeutung man noch nicht ergründen bzw. gelernt hat, ist unmöglich."[1] (Jan Walburg)

Therefore, now it is essential to find the connections, regulatory mechanisms and functions of the "words" (genes). It is not possible to unveil all the functions of a whole genome of an organism in one master thesis, thus the author will concentrate to unravel circadian regulated genes and to estimate the influence of light on the gene expression of rhythmically expressed genes. The work also encompasses the analysis of different plant tissues to test the assumption that the circadian clock is organ specific. On a rudimentary and purely technical level the impact of replicates and randomization techniques onto the significance of the results will be examined.

[1] For reading a book that is written in a language of which one knows the alphabet, but whose words and meaning one does not fathom or have learnt yet, is impossible. [translation, WW]

2. Theoretical principles

2.1 Circadian Clock

The plant circadian clock and its specific components were investigated predominantly in *Arabidopsis thaliana*. The examination of internal biological rhythms and their underlying concepts is done by the scientists in chronobiology. In this field of science various species and different levels of biology, starting with the elucidation of molecular mechanism in protazoa up to the complex study of psychological phenomena in humans are considered. The constant interdisciplinary exchange of information between the scientists is unique in natural sciences. The popularity arose over time and once again it was a Frenchman who gave the necessary impulse. In 1729, the opening of Mimosa leaves during the morning and its closure with the beginning of the night attracted the attention of the Astronomer Jean Jacques d'Ortous de Mairan [13]. It seems to be obvious, that this rhythm was associated and influenced by the prevalent light conditions. To test this theory he exposed a Mimosa plant to a dark cupboard and observed that the leaf movement persisted. This was the first experimental evidence that the internal rhythm controls diurnal rhythm in plants without external stimuli. Nevertheless there were different opinions on the origin of these internal rhythms. Some scientists believed that a yet unknown variable in the environment that is correlated with time of day might be instrumental in eliciting biological periodicity. The other assumption was that the rhythm is generated within the cells of the plant. In the year 1793 the German philosopher, mathematician and physical experimentalist Georg Christoph Lichtenberg postulated the theory of an internal clock in humans and pointed out, that such a mechanism assumed internal clock-like structures [46]. Not everybody agreed to this idea and a lot of research time was spend to hunt for the unknown correlate of the earth's axial rotation (factor 'X') that drove daily cycles [13]. If the scientists had better interpreted the observations of the Swiss botany Augustin de Candolle from 1832 they could have saved themselves the effort. He noted that the daily rhythm of leaf movements of Mimosa plants housed in continuous illumination had a cycle length up to 2 h shorter than the exact 24 h periodicity seen in natural daylight [73]. As a

side- effect the plant is desynchronized of the daily light/dark cycle in nature. This proved that the organism regulates the periodical opening and closure of the leaves itself. At the beginning of the twentieth century the experiments from *Anthonia Kleinhoonte* [43] and *Erwin Bünning* [7] confirmed the assumption and closed the controversy on the search of factor 'X'.

The initial research in the field of chronobiology was fully dominated by the studies on the 'sleep movement' of plants. If the analysis had been done for humans the observations of de Candolle would have been confirmed earlier. As happened by the experiments of the German biologist and behavioral physiologist Jürgen Aschoff. From 1967 to 1979, as a director at the Max Planck Institute for Behavioral Physiology in Seewiesen, he investigated the influence of external stimuli on the endogenous circadian system in humans. Within a few test series he verified the observations by de Candolle. In the following years the institute became the mecca of chronobiology.

2.2 What is a circadian rhythm?

The denotation “circadian” was mentioned for the first time by Franz Halberg in the year 1959. The term comes from the Latin *circa*, meaning “around” or “approximately” and *diem* or *dies*, meaning “day”. It should emphasize that 'about a day' rather than precisely 24 h is the true hallmark of these periodicities [13]. So, one always speaks of a circadian rhythm, if the period length is between 22 and 25 hours. For shorter periods the term 'ultradian' is used and in the case of longer periods one speaks about 'infradian' rhythms. Processes are defined as outputs of the circadian clock rather than mere responses to environmental cues if they meet the following criteria. First, circadian rhythms persist with approximately (but never exactly) 24-hour periodicity after an organism is transferred from an environment that varies according to the time of day (entraining conditions) to a constant environment (free-running conditions). Second, the time of onset of these rhythms can be reset by appropriate environmental cues, such as changes in light or temperature levels. Finally, circadian rhythms are temperature

compensated; that is, they occur with approximately the same periodicity across a wide range of temperatures [28]. Most biological processes show a Q_{10}-value of 2-3, that means that the reaction velocity is doubled/tripled at a temperature change of 10°C [39]. In contrast, the circadian clock of *Arabidopsis thaliana* has a Q_{10}-value of 1.0- 1.1 over a range of 20°C [74]. Physiological experiments performed in a variety of model organisms, including plants, animals, fungi, and cyanobacteria, revealed fundamental commonalities between the circadian systems of these diverse species [28].

2.3 How are circadian rhythms generated?

By taking advantage of the physical theory of oscillators, Aschoff was able to predict the behavior of circadian systems and he declaimed the idea of a natural biological oscillator. The first very simplified model which assumes that the input signals entrain the core oscillator which generates the rhythmic activity of the internal clock is shown in Figure 2.1 A. However, this linear progression is an oversimplification, because many components of the input pathways are themselves outputs of the clock and rhythmic outputs from the clock may feed back to affect the functioning of the core oscillator [21]. Figure 2.1 B takes this consideration into account. While concentrating on the model it is pertinent to ask why circadian clocks appear to have converged on such complex architectures. The answer is simple: the complexity provides the clock with stability and protection against stochastic perturbations [21]. Additionally every single gene has its own specific time point when it reaches its maximal expression in the course of the day. The complex structure of the clock ensures the required flexibility to coordinate the different cellular- mediated processes. The circadian system is highly adaptable and dynamic. Exogenous signals influence and even more importantly synchronize the internal clock with the surrounding environment. Such influencing variables like light or temperature cycles were called 'Zeitgeber' by Aschoff for the first time and the term was accepted in the English- speaking scientific community.

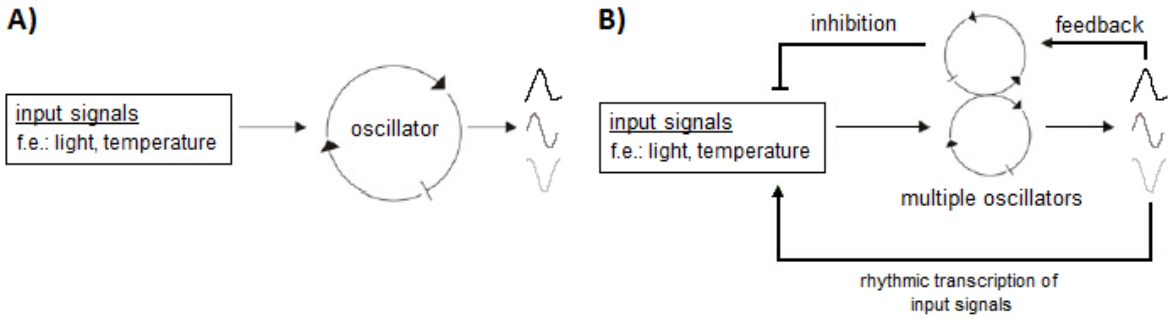

Figure 2.1: Schematic representation of circadian clock structures
A) Simple model consisting of input signal, oscillator and rhythmic output B) Elaborated clock model taking into account additionally multiple oscillators and outputs which feed back into the central oscillator. Arrows are positive arms and perpendicular lines represent negative arms of the pathway [21]

By circadian convention, the time of onset of a signal that resets the clock is defined as Zeitgeber time 0, abbreviated ZT0 [28]. This definition is especially used for time scales under light/dark cycles (LD conditions). Whereas under constant light conditions the term 'circadian time', abbreviated CT, is commonly used [82]. ZT0 indicates the time point when light is turned on and ZT12 when light is turned off. CT0 in contrast marks the subjective dawn and CT12 the subjective dusk. Why is the term 'subjective' used? Under constant light conditions no external time cues exist and therefore it could only be a subjective time point [82]. The influence of the Zeitgeber signal on the phase of the rhythm varies with the time of day. Depending on which time point the signal is given a phase shift is possible as well as cells become desynchronized. There is also the possibility that nothing changes [65].

2.4 How are circadian rhythms determined?

The output of the circadian clock often takes the form of sinusoidal waves that can be described by mathematical terms such as period, phase and amplitude [28]. Figure 2.2 shows an idealized clock output in light/dark cycles with important parameters highlighted.

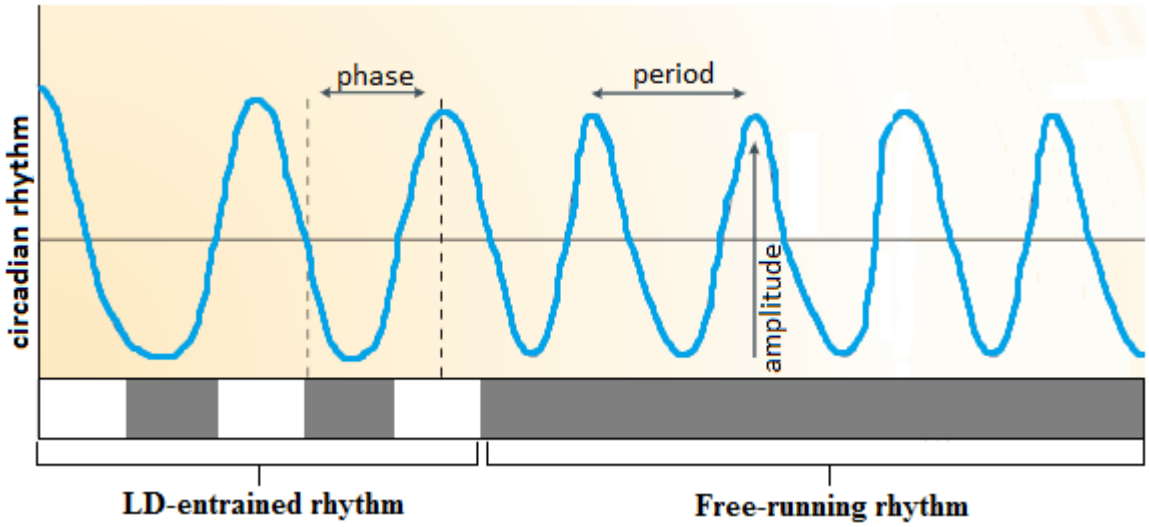

Figure 2.2: Parameters of circadian rhythm
An idealized circadian rhythm is depicted. The difference in the level between peak and trough values is the *amplitude* of the rhythm. The timing of a reference point in the cycle (e.g., the peak) relative to a fixed event (e.g., beginning of the night phase) is the *phase*. The time interval between phase reference points (e.g., two peaks) is called the *period*. [http://bioinformatics.cau.edu.cn/ARSER/ismb2010_rdyang.pdf]

There are many different methods to calculate the parameters which are necessary to model circadian rhythms. The most challenging part is the determination of the accurate period. It has to be taken into account that the expression of genes is a continuous process, which is transformed into a discrete time series during the experimental procedure. This results in an unavoidable loss of information and the reconstruction is a big challenge. In the field of information theory the number of necessary sampling points could be calculated to perfectly reconstruct the original function. The theorems state a lot of sample- rate criterions and therefore in practice the perfect reconstruction is replaced by mathematical approximations, because under different circumstances or conditions these criterions are not satisfied. Nevertheless the ideas, especially the one that always equidistant sample rates are used influenced the sampling methods in time- series experiments.

In the field of chronobiology, microarray experiments with an equidistant interval of time points are commonly used. Circadian microarray experiments are usually designed to collect data every 4 h over a course of 48 h, generating expression profiles with 12 or 13

time-points [86]. Figure 2.3 shows a schematic design of such a typical microarray experiment.

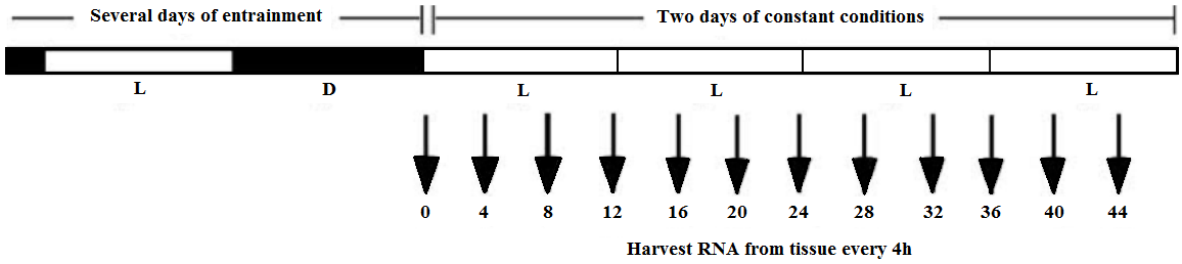

Figure 2.3: Schematic representation of a circadian microarray experiment
In the beginning of the experiment the plants are entrained in light/dark cycles. After several days they are transformed to constant light conditions and samples are collected every 4 hours.
[http://bioinformatics.cau.edu.cn/ARSER/ismb2010_rdyang.pdf]

The number of data points is limited due to budget constraints and working time. Therefore the sampling rate is relative low and it increases the difficulties to generate statistically significant results. Furthermore most of the commonly used methods are not feasible for such short time-series.

Two of the oldest and multiple applied methods are the Fisher's G-test and the COSOPT algorithms. COSOPT is time-domain algorithm which matches gene expression profiles to a predefined selection of 101 cosine curves of varying phases. The quality of the pairs is specified by a multiple-measures corrected β (MMC-β) value. A very small MMC-β value indicates a nearly perfect match between the curve and the experimental data. All genes with a threshold value smaller or equal to 0.05 are classified as rhythmically expressed. The big advantage of the algorithm is its efficiency for short time- series, but it's noise-sensitive and model-dependent. The alternative is a frequency-domain algorithm like the Fisher's G-test. This test generates a periodogram of experimental time-series and tests the significance of the dominant frequency [86]. Classification takes place according to the g-statistics. A high value leads to a rejection of the null-hypothesis, which assumes a random event. This method is model-independent, but performed worse on short time-series. A lot of other techniques for determining rhythmically expressed genes were developed like in *Hughes et al.* [34], *Luang and Li* [50], *Lu et al.* [48], *Wichert et al.* [84] and

Zhao et al. [90]. These algorithms are only slight variations of the described methods and will therefore not be further explained. Recently, *Yang and Su* [86] introduced an algorithm named ARSER, which combines time domain and frequency domain analysis. ARSER is optimized for time- series with 12-13 time- points. The algorithm predicts the periodicity of a gene expression profile by employing autoregressive spectral estimation. In the next step harmonic regression is used to model the rhythmic pattern. ARSER performed best on periodic and non-periodic. For that reason the algorithm was chosen for the analysis and is described more precisely in chapter 3.4.1 later on.

2.5 Circadian internal synchronization and desynchronization

The system of endogenous circadian oscillators has been investigated in organisms ranging from unicellular algae to man [57]. Nevertheless it is still a matter of debate whether one master oscillator controls all circadian phenomena or several Zeitgebers are responsible to synchronize each circadian rhythm [33]. Now considerable evidences indicat the existence of a multioscillator system which generates the different rhythms autonomously in a single organism [57]. All living organisms are organized temporally to insure that there is “internal synchronization” between the myriad biochemical and physiological systems in the body [78, 81]. Internal synchronization demands that within an organism there must be only one oscillator or “clock” on which all endogenous circadian rhythms are passively dependent, or, if there is more than one oscillator, then the various oscillators must be normally synchronized with one another [57]. Thus the only way to proof that circadian rhythms are regulated by different autonomous oscillators is to demonstrate desynchronization of the rhythms under constant external conditions. In the absence of external cues the rhythms are said to be “free-running”, meaning that no synchronization by any cyclic change in the physical environment occurs, with a period close to but not exactly 24 hours [81]. This period of an endogenous rhythm when light and temperature are held constant was called “natural period” by *Pittendrigh and Bruce* [61]. However, if the rhythms are controlled by the same oscillator, the periods

have to be identical. The term “internal desynchronization” is used to describe a state where different oscillating variables within an organism demonstrate different periods and therefore constantly changing phase relationships [57]. It has to be mentioned that a lack of internal desynchronization does not necessarily deny the existence of a multioscillator system [33]. *Roenneberg and Morse* [67] were one of the first who could proof the existence of more than one oscillator within a single cell. They investigated three different rhythms in *Gonyaulax* and under certain experimental conditions these rhythms run independently. Indicating that each rhythm must be controlled by its own distinct oscillator [67]. In plants multiple circadian rhythms occur in many species and *Hennessey and Field* [31] could show that they have different free-running periods, indicating anatomically distinct oscillators. The transfer from light/dark cycles to constant light condition may lead to “transients”, so that the phase is not reset immediately but comes to a stable position after several cycles [29]. It should be referred to “transient internal desynchronization” if the internal desynchronization is observed between two stable synchronized states [57]. The time depends on the organism and may be related to the relative complexity of the organisms [29]. If two or more rhythmic variables are shown to free-run with different frequencies for a sufficient length of time the term “steady-state internal desynchronization” is used [57]. *Roenneberg and Morse* further suggest that chloroplasts, which evolved from an endosymbiotic cyanobacterium, had contributed their own oscillator to the circadian systems of eukaryotic cells. It could be observed that not only the gene expression in the nuclear genome but also in the chloroplast genome is circadian regulated. *Roenneberg and Morse* formed their assumption based on the experiments in the giant alga *Acetabularia*, which demonstrated that the chloroplast photosynthetic rhythm persist even in an enucleated cell [67]. In contrast, *Matsuo et al* [53] provided direct evidence that the circadian period of chloroplast gene expression rhythms is determined by the nucleus-encoded circadian oscillator. However, they could not exclude the possibility of the existence of a chloroplast specific clock.

3. Material and Methods

3.1 Microarray experiments

Three different datasets were used to analyze the gene expression in two organisms, tissues and conditions. To analyze the gene expression in *Arabidopsis thaliana* we used the diurnal time-series of 3 biological replicates generated by *Bläsing et al.* [5]. They harvested plants every 4 hours at 6 time-points beginning with the end of the night. The details are listed in the appendix. To investigate the influence of elongated day length on the oscillating genes the dataset from *Kim et al*. [42] was used. In their experiments *N. attenuata* plants were grown in 16 h light/8 h dark cycle. The data were collected every 4 h and lead to a time series gene expression data set with 6 time-points. To examine local effects, different tissues of three biological replicates were collected. The details are also listed in the appendix. The last dataset contained gene expression data from *N. attenuate* grown under light/dark rhythm and constant light conditions. The design is shown in figure 3.1.

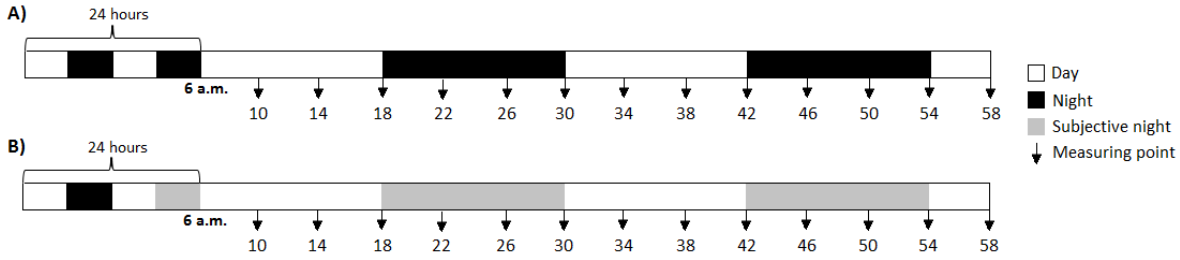

Figure 3.1: Design of the microarray experiment for Nicotiana attenuata under LD and LL conditions
A) Plants were grown in a 12/12 h light/dark cycles and harvested every 4 h for two days B) Plants were entrained under day/night rhythm for 5 weeks and then exposed to constant light conditions. Sampling was started at 28 h after constant light exposure.

One part of plants was grown in 12 h light and 12 h dark cycles and harvested every 4 hours over two days, which leads to a time- series of 13 time- points. The rest of plants were first entrained in a 12 h light and 12 h dark cycles for 5 weeks and exposed to the constant light conditions. The sampling was started at 28 hours after transferring plants and the same rate was used as for plants grown under LD conditions.

3.2 Normalization

The present data emerged from a number of microarray experiments and to compare the different probes a normalization method had to be applied. This step is necessary to reduce the bias which could distort the results. Before the real normalization process is performed the dataset had to be checked for possible errors. Potential sources of such errors are the preparation of mRNA, fluorescent labeling, hybridization procedure or scanning of the array, *Victor et al.* [83]. In the field of explorative data analysis the results of microarray experiments are visualized by boxplots, histograms or quantile-quantile- plots. The graphics provide the first overview so that conclusions about the variance and distribution of the data can be drawn. Outliers are any data not included in the box, and were plotted with a small circle. The highlighted boxes in the diagram visualize the area which contains 50% of the average data. The spacing between the bottom and top of the box indicates the degree of dispersion in the data, because the length of the box corresponds to the interquartile range. Therefore the first and third quartiles are marked by the upper and lower band respectively. Additionally the horizontal line inside the box marks the second quartile, the median. That means 50% of the data are above the band and 50% below. From the position of the median conclusions about the skewness in the data can be drawn.

Until now no “gold standard” exists. In the literature various techniques were tested and their shortcomings listed, *Boes et al.*[6]. Especially at the beginning of microarray technology the use of so- called 'housekeeping' genes was one of the favorite methods. The assumption of this method is that there are some genes which keep their expression constant over time and are independent of external stimuli. The genes are hybridized as zero stock checks. This kind of normalization technique calls for excellent knowledge about the gene expression of the currently investigated organism. Only in rare cases the necessary information is available and verified. Besides the whole circumstances are under consideration, because it cannot be ruled out that the selected genes change their expression under special conditions. The average intensity of the control genes is

calculated and that followed a subtraction of this value from all other values in the original dataset. As the number of 'housekeeping' genes is low compared to the whole amount of genes on a chip the intensity of all genes is increasingly used to avoid incorrect scaling. The use of ‘housekeeping’ genes is based on the assumption that the majority of examined genes are not relevant as far as this investigation is concerned and therefore only a small number of genes decisively change their expression. Different measures of location are applied. In most cases the 0.75- quantile of all genes is taken and every expression value is divided by this value. Another popular value is the second quartile, the median. In this work the 0.75 quantile normalization was used and the data was log2- transformed.

3.3 Simulated time series

We used a randomization procedure to combine the expression value of the pooled biological replicates to new time-series. The underlying assumption was that the amplification of the pooling effect reduces the number of expression profiles incorrectly classified as periodic (false positives). For the experimental data we are not able to know which genes are true (expression profiles correctly classified as periodic) and false positives and the literature does not provide this information for other plant species as well. To test our methodology we generated artificial time series data with known numbers of true and false positives. Each time-series contains expression values of 30000 genes and consists of an alternating ratio of periodic and non-periodic expression profiles. For Arabidopsis a number between 6% and 15% of circadian regulated genes within microarray time course data is estimated [56]. Therefore we decided to generate datasets with ratios of 5%, 10% and 15% of periodic time-series data. Stationary and non-stationary models were used to simulate periodic patterns as proposed by *Yang and Su* [86]. The stationary model is defined by

$$x_t = SNR \cdot 2cos\left(\frac{2\pi}{\tau}t - \varphi\right) + \varepsilon_t \quad (1)$$

and the non-stationary is defined by

$$x_t = 500 \cdot e^{-0.01t} + SNR \cdot 100 \cdot e^{-0.01t} \cdot cos\left(\frac{2\pi}{\tau}t - \varphi\right) + \varepsilon_t \quad (2)$$

where SNR is signal-to-noise ratio; τ is period; φ is phase; and ε_t is (μ=0, σ=1) normally distributed noise terms. Additionally, a combined model of cosine and sinus functions described by *Wichert et al.* [84] was used to generate another periodic expression profile

$$x_t = cos\left(\frac{2\pi}{\tau}t\right) + sin\left(\frac{2\pi}{\tau}t\right) + \varepsilon_t \quad (3)$$

As common procedure we used normally distributed white noise to simulate non-periodic time series. Based on the results of *Futschik and Herzel* [20] we also included autoregressive (AR) processes of order one as a model for non-periodic expression profiles. An overview of the datasets is given in table 1.

Table 1: Overview of simulated datasets given the exact number of periodic and non-periodic patterns as well as the absolute number of expression profiles for the different models

model	**dataset1**		**dataset2**		**dataset3**	
Stationary	1500		1000		500	
Non-Stationary	1500		1000		500	
Combined	1500		1000		500	
White noise	12750		13500		14250	
AR	12750		13500		14250	
Periodic	4500	15%	3000	10%	1500	5%
Non-periodic	25500	85%	27000	90%	28500	95%

For each dataset we generated 3 time-series which should perform as biological replicates. The expression values of these time-series were randomly combined to generate 30 time-series consisting of 30000 genes with 12 time-points over a range of two days. The details of this randomization procedure are described later on.

3.4 Detection of rhythmically expressed genes

3.4.1 ARSER

We used the ARSER algorithm from *Yang und Su* [86] to detect the oscillating genes within the time course series. The algorithm combines time- domain and frequency-domain analyses to address such rhythmically expressed genes based on their expression profiles within microarray data. Via harmonic regression the algorithm is able to model the rhythm using four parameters: period (duration of one complete cycle), the mean level (the mid-value of the time-series), the amplitude (half the distance from the peak to the trough of the fitted cosine, indicating the predictable rise and fall around the mean level) and the phase (the location in time of the peak of the curve). These values and other analytical results like statistic validation are returned in an output file from the algorithm [].

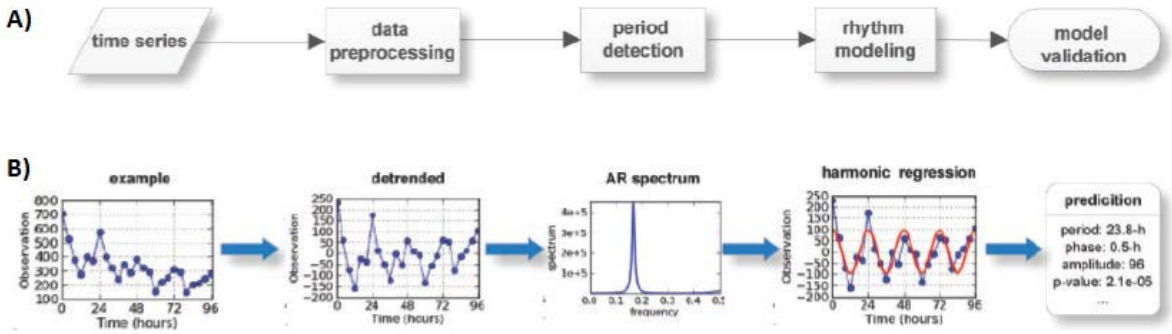

Figure 3.2 : Schematic representation of the ARSER algorithm and a case study.
A) Analysis flowchart. First, data pre-processing by linear trend removal (detrending), then period detection by searching peaks from the AR spectrum. With the periods derived from the AR spectrum, harmonic regression is carried out to model circadian rhythms by fitting the detrended time-series with trigonometric functions. B) An example of rhythmicity analysis by ARSER [86].

The algorithm is optimized for time course series with 12-13 data points over a range of 48 hour. ARSER performs a data preprocessing strategy called *detrending* that removes any linear trend from the time-series so that we can obtain a stationary process to search for cycles. Detrending is carried out by ordinary least squares (OLS) [86]. This step removes a long time trend from a time course series to prevent a distortion of the results. A time

course series is called stationary if statistical properties stays constant over time. Afterwards autoregressive (AR) spectral analysis is carried out to search for circadian rhythms and to determine the period. With the periods obtained from AR spectral analysis, ARSER employs harmonic regression to model the cyclic components in time-series. Finally, when analyzing microarray data, false discovery rate (FDR) *q*-values were calculated for multiple comparisons [86]. A schematic overview of the methodology is shown in figure 3.2.

The most important step is the determination of the accurate period length, because the period can differ from the assumed length of 24 hours. The algorithm takes a range from 20 to 28 h into account. The AR spectral analysis calculates the power spectral density of the time-series in the frequency domain. If there are cycles of circadian period length in the time-series, the AR spectral density curve (equation 2) will show peaks at each associated frequency [86]. At the beginning an AR model of order p is generated to fit the time-series using the following equation:

$$x_t = \sum_{i=1}^{p} \alpha_i x_{t-i} + \varepsilon_t \tag{1}$$

where ε_t is white noise and α_i are model parameters with $\alpha_{ip} \neq 0$ for an order p process. AR coefficients are generally estimated by three methods: Yule-Walker method, maximum likelihood estimation and Burg algorithm [86]. To fit the model to the experimental data the order p is set to $24/\Delta$ and the coefficients are calculated using all three methods mentioned above and following equation:

$$p_x(\omega) = \frac{\sigma_\varepsilon^2}{|1 + \sum_{k=1}^{p} \alpha_k e^{-i\omega k}|^2} \quad 0 \leq \omega < \pi \tag{2}$$

where σ^2_ε is the variance of white noise; α_k are parameters defined in Equation (1). All periods within the spectrum were selected. The determined period of the time- series serves as an input for the following harmonic regression:

$$x_t = \mu + \sum_{i=1}^{n} \beta_i \cos(2\pi f_i t + \phi_i) + \varepsilon_t \quad (3)$$

where x_t is the observed value at time t, μ is the mean level of the time- series; β_i is the amplitude of the waveform; ϕ_i is the phase, or location of peaks relative to time zero; ε_t are residuals that are unrelated to the fitted cycles; and t are the sampling time- points. For spectral analyses the period is always predefined and equation 3 can be simplified to a multiple linear regression model:

$$x_t = \mu + \sum_{i=1}^{n} \{p_i \cos(2\pi f_i t) + q_i \sin(2\pi f_i t)\} + \varepsilon_t \quad (4)$$

where $p_i = \beta_i \cos \phi_i$ and $q_i = -\beta_i \sin \phi_i$. The unknown parameters p_i, q_i and μ can be estimated by OLS method. Then the amplitude β_i and phase ϕ_i are obtained by $\beta_i = \sqrt{p_i^2 + q_i^2}$ and $\tan \phi_i = -q_i / p_i$ .To validate the results different statistical properties were calculated. The correlation coefficient is calculated to find the cosine curve which fits best to the expression profiles. Thereby various cosine curves with different parameters were matched to the experimental data. The function with the highest correlation coefficient was selected. The results were statistically validated by calculating the p-value. In this regard the p- value indicates the probability that a randomly chosen expression profile has a correlation coefficient higher than a predefined threshold value. Small p- values mean it's unlikely that randomly chosen expression profiles receive a correlation coefficient higher than the threshold and lead to a rejection of the null hypothesis [82]. By doing so it's ensured that the error rate stays as low as possible. Typically, data from microarray experiments provide information about many genes simultaneously. By analyzing the data one faces the problem of multiple testing and the need to adjust the p- value. There are different methods to do so and here the method from *Storey and Tibshirani* [76] was used. The most important fact is that their approach discriminates explicitly between the concept of FDR and false positive rate. The false positive rate refers to the probability of not significant results falsely classified as significant, whereas the FDR states that

significant features are truly not significant. For example, a false positive rate of 5% means that on average 5% of the truly null features in the study will be called significant. A FDR of 5% means that among all features called significant, 5% of these are truly null on average [76]. Valid conclusions about the false positive rate and the FDR can be made by stating the feature specific p- and q- values. Whereas the p- value is a measure of significance in terms of the false positive rate, the q- value is a measure in terms of the FDR [76]. A p- value of 0.05 accords to a false positive rate of 5% and doesn't tell much about the content of the features actually called significant. Better to use the q- value which provides a measure among the features called significant. The FDR is a sensible measure of the balance between the number of true positives and false positives. The ARSER algorithm calculates the q- value, as well as the p- value, for every gene within a dataset. Genes with a q- value below 0.05 are classified as rhythmically expressed ones.

3.4.2 Haystack

HAYSTACK is designed to find periodic patterns in any large-scale dataset representing at least three data points. The Web version is available at http://haystack.mocklerlab.org/. This algorithm compares the experimentally recorded gene expression profiles with predefined cycling patterns and groups genes whose expression profiles match the same or similar patterns. Different cutoffs are used to detect circadian oscillation. The most important parameter is the correlation coefficient. The higher the value the higher the degree of correlation between the experimental data and the different models. A coefficient of +1 indicates perfect positive correlation. Another cutoff value which is taken into account is the fold change. This value describes how much the initial value differs from the finale one. To achieve statistical significance the p- value is also calculated. Periodic patterns within the random datasets were identified using default parameters except that a fold cutoff of 1.0 instead of 2.0 was applied. In the case of time course series the most biologically relevant information are whether a transcript cycles and if so, the timing or phase of its maximum expression over the day. The common

algorithms which use the same principle search for significant cross-correlations with sine or cosine waves. What is special about HAYSTACK is its possibility to search for at least six different patterns, including “asymmetric”, “rigid”, “spike”, “cosine”, “sine”, and “box-like” patterns. The most successful models in identifying rhythmically expressed genes are “cosine” and “spike”. This means that these patterns show the highest correlation coefficient. A set of model profiles was downloaded from the HAYSTACK web site mentioned above. This dataset included all the patterns listed earlier except the “asymmetric” one. Each periodic pattern contained 24 samples so that a total of 120 time-series were available. Each series possessed 12 time-points that represent two circadian cycles obtained at 4 h sampling intervals.

3.5 Estimation of molecular peaking time

To estimate the peak time of cycling genes the method proposed by *Ueda et al.* [82] modified for our requirements. *Ueda et al.* used an algorithm like COSOPT. 24-h period cosine curves with different peak times at 10-min intervals are predefined as test patterns. COSOPT calculates the correlation coefficient between the 12-point time course of each gene and the test patterns. The best-fitted cosine curve was selected and they defined the molecular peak time for each cycling gene as the peak time of that curve. We instead used the parameters returned by ARSER to model the rhythm for every gene which was classified as rhythmically expressed. This modeling step leads automatically to the best fitting curve for the expression profile of each gene which seems to be circadian regulated. Afterwards the maximum, in this case called ‘peak time’, of the curve was calculated and assigned to the gene.

3.6 Determination of ZT-, CT- and similarity groups

The genes which are classified as rhythmically expressed often vary a lot concerning the parameter values. There is the possibility to cluster the expression profiles based on their pattern. This method leads to a number of classes with different sizes and

comparisons between the groups are difficult. Such a grouping is not helpful to compare expression profiles from genes under different conditions. It would be preferable to use an equidistant scale to group the genes like it was proposed by *Ueda et al.* for LL conditions. Here the genes are divided into various CT groups based on their molecular peaking time. The parameters returned from the ARSER algorithm were used to model the expression profile of every gene which was classified as rhythmically expressed. Based on these functions the molecular peaking time was estimated. The genes were divided into 24 CT groups with fixed boarders. CT0 contains genes with a peaking time between 23.5 and 0.5. Genes with a peaking time in the interval (0.5, 1.5] were grouped to CT1 and so on. The upper limit belonged to the currently regarded group and the lower limit to the previous group. For example a gene that showed maximal expression at 0.5 belonged to CT0. *Kerwin et al.* [41] referred to this method in their paper about single time-point analysis and were able to achieve good results to uncover natural variations concerning the function of the circadian clock in *A. thaliana*. By convention, a similar equidistant scale is used to group genes under LD conditions. There are 24 zeitgeber times (ZT) hours per day and lights on (6 a.m.) and off (6 p.m.) are designated ZT 0 and 12 [32]. The same borders were used as for the CT group determination, so ZT0 contains genes with a peaking time between 23.5 and 0.5 and so on. Once again the upper limit belonged to the currently regarded group and the lower limit to the previous group.

3.7 Cluster analysis

Based on the proposed principle from *Ernst et al*. [18] we developed an algorithm for clustering short time series expression data. The algorithm works by assigning genes to a predefined set of model profiles that take into account that genes can be assigned to specific time bins based upon coincidence of peak expression within a circadian period. 24 CT groups were defined as explained in the previous section. Each group had a single peak of expression per day. We used the artificially created datasets for the LD conditions to define the parameters of the profiles. The genes of these time course series were already

classified as rhythmically expressed ones or not and if so assigned to the right CT group. Therefore we calculated the median of the phase, amplitude and period of every CT group. We randomly selected an artificial dataset to determine the parameter values and to design the model profiles for the time bins. The profiles form the basis for the clustering process and were mapped to the expression values by calculating the correlation coefficient *c*(profile, gene expression values). The algorithm always returns the three CT groups with the highest correlation coefficient for every gene. We assumed that a high correlation between the model profiles and the expression values indicates a periodic pattern within the time course. To distinguish between rhythmically expressed genes and such which only matched loosely to the profiles we set a threshold of 0.9 for the correlation coefficient.

4. Results

4.1 Normalization

In the field of explorative data analysis the results of microarray experiments are visualized by boxplots, histograms or quantile- quantile- plots. The graphics provide the first overview so that conclusions about the variance and distribution of the data can be drawn. Figure 4.1 shows the boxplots of logarithmic but not normalized expression values from three examples in a single time-point (ZT4) under LD (A) and constant light (LL) (B) conditions. From the position of the median conclusions about the skewness in the data can be drawn. The boxplot reveals that the distribution of the replicates is skewed to the right. To allow comparison among replicates the data has to be normalized. In most cases the 0.75- quantile of all genes is taken and every expression value is divided by this value. Another popular value is the second quartile, the median. In this work the 0.75 quantile normalization was used and the data was log2-transformed. The effect of the normalization procedure on the distribution in the data demonstrates in Figure 4.2. Once again the boxplots of the logarithmized expression values from three replicates in a single time-point (ZT4) under LD (A) and LL (B) conditions are shown, but now the values were also normalized. Identifiable by the fact that the upper band of the boxes are at the same height. Differences between the three replicates are also highlighted by normalization. The distribution of the outliners is also interesting to note. For replicate 1 the outliners were highly concentrated and were closely interrelated. In contrast replicate 2 and 3 had a broad spectrum of outliners.

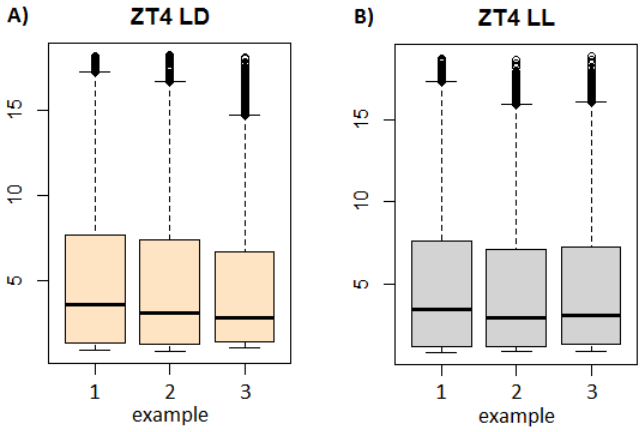

Figure 4.1: Representative boxplots for ZT4
A) boxplots of logarithmized but not normalized expression values from three under LD condition B) boxplots of logarithmized but not normalized expression values from three under LL condition

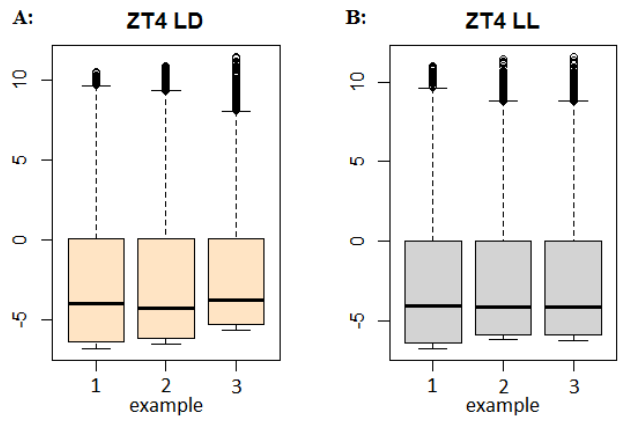

Figure 4.2: Representative boxplots for ZT4
A) boxplots of logarithmized and normalized expression values from three under LD condition B) boxplots of logarithmized and normalized expression values from three under LL condition

4.2 Experimental microarray data

To unveil the diurnal rhythm of *N. attenuata* transcripts, we used kinetic microarray in which three biological replicates were available. Leaf samples were collected every 4 h for two days to extract RNA and subsequent were used for microarray analysis. The individual plants do not deliver enough material for a whole time course, therefore the samples were pooled. This method combines the material from different individuals in a pooled sample before labeling and hybridization are done [40]. So the biological replicates are more or less randomly generated from the whole RNA pool. Each replicate was hybridized on different array to reduce the bias technical replicates. Every plant has its own genetic characteristics although they all belong to the same species and are inbred lines. The boxplot of the unnormalized data (Figure 4.1) illustrates the differences among biological replicates. Especially the distribution and amount of outliers differ a lot. The applied 0.75-quantile normalization reduced the noise and allows a comparison among the replicates. The microarray data was analyzed with ARSER, an algorithm that combines time- domain and frequency- domain analysis to detect periodicity in gene expression profiles and is described more precisely in chapter 3.4.1. ARSER returns four parameters to describe the rhythmic patterns: period, phase, amplitude and mean level, and measures

the multiple testing significance by FDR *q*-value [86]. All genes with a *q*-value below 0.05 were classified as oscillating genes and selected for further analysis. The absolute and relative frequency of oscillating genes varies a lot between the replicates as can be seen in Figure 4.3. For replicates 2 and 3 from plants grown in 12 h light and 12 h dark cycles nearly the same amount of rhythmically expressed genes was predicted, whereas ARSER detects almost twice as much oscillating genes for the first replicate (Figure 4.3A). The relative frequency of rhythmic patterns in gene expression profiles of plants grown under diurnal light conditions alternates between 28% and 66%. Although the data was normalized a large variance between the replicates could be observed under LD conditions. The situation becomes even clearer when we focus on the data from LL conditions (Figure 4.3B). Replicate 2 differed dramatically from the other two replicates. Even more surprising was that the number of oscillating genes in replicate 2 was higher than the numbers for replicate 2 and 3 in LD conditions. The number of rhythmically expressed genes in LL conditions was fewer than that in LD conditions.

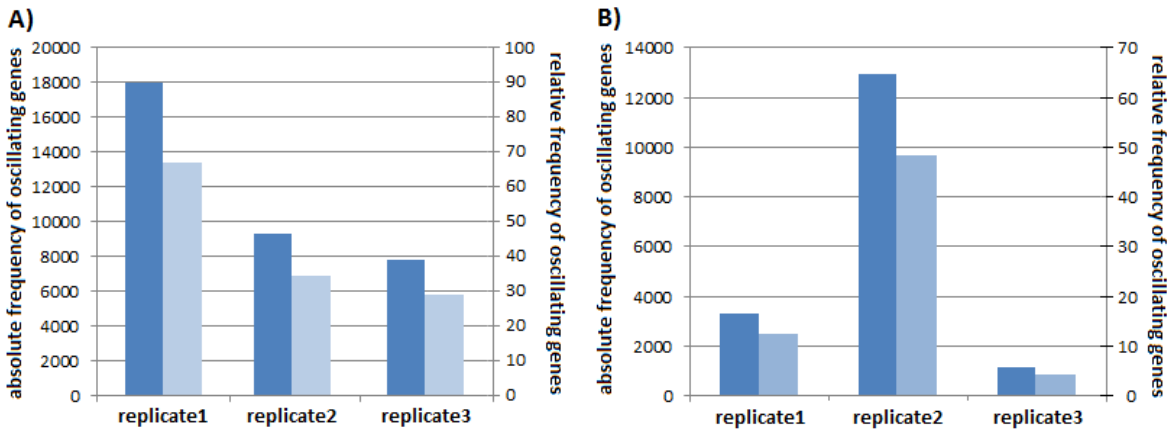

Figure 4.3 Absolute and relative frequency of oscillating genes

A) Absolute (dark-blue) and relative (light-blue) frequency of oscillating genes for three biological replicates (replicate 1-3) under LD condition. B) Absolute (dark-blue) and relative (light-blue) frequency of oscillating genes for three biological replicates (replicate 1-3) under constant light condition. The number of rhythmically expressed genes was predicted by ARSER with a stringency threshold (q-value) set to 0.05.

For Arabidopsis microarray time course data a number between 6% and 15% of circadian regulated genes is predicted [56]. Therefore the results of replicate 1 and 3 which indicate a relative frequency of 12% and 4% respectively seems to be the most reliable. Nevertheless the second replicate may not just be ignored especially because the expression values of the replicates 2 and 3 in constant light condition differ that much as one can see in the boxplots (Figure 4.2). The returned parameters from the ARSER algorithm were used to model the rhythmic expression profiles of the oscillating genes and the maximum of the curve was calculated. Based on the peaking time the rhythmically expressed genes were assigned to various ZT groups for LD condition and CT groups for LL condition. The whole procedure is described in chapter 3.5 and 3.6. Figure 4.4 shows the CT group distribution of the three biological replicates (relative frequency).

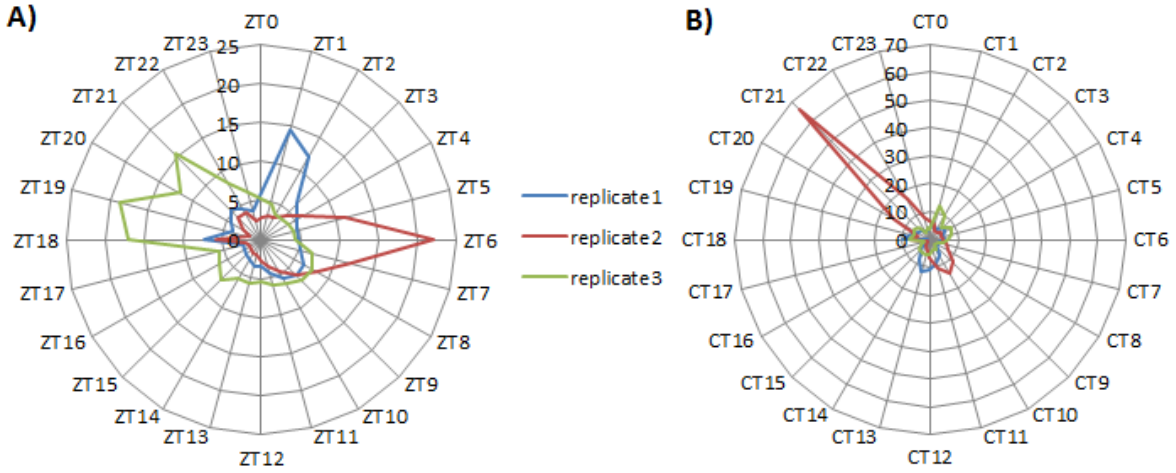

Figure 4.4 ZT (A) and CT (B) group distribution for both light conditions

A) Relative number of genes per CT interval for three biological replicates under LD condition B) Relative number of genes per CT interval for three biological replicates under LL condition

Each replicate shows a specific distribution pattern and the overlap is really small for both light conditions. Most of the genes of replicate 1 were assigned to ZT1 and ZT2 that means the majority of genes reached their maximal expression right after dawn, whereas most of the genes of replicate 2 peaked around midday (ZT6). In contrast, most of the genes of replicate 3 reached their maximal expression in the middle of the night.

If we concentrate on the LL condition data we could observe a similar effect. As can be seen in Figure 4.3B the highest number of oscillating genes was predicted for the second replicate and over 50% of the genes were assigned to CT21. Such a huge amount could not be observed for any other replicate or conditions. This explains why the CT group distribution of the other two replicates was not immediately ascertainable. Therefore Figure 4.5 shows the distribution of replicate 1 and 3 in more detail. From it, we see that the replicates have nearly an opposite distribution pattern. Most of the genes of replicate 1 were assigned to CT1, whereas the oscillating genes of replicate 3 mostly peaked at dusk. To sum it up, on the basis of the presented data we can say that the gene expression profiles of the three replicates seems to vary a lot and as a result the number of predicted oscillating genes as well as the ZT-/CT- group distribution differs significantly between the three replicates. Nevertheless it is interesting to know, whether the lists of rhythmically expressed genes for each replicate had some genes in common. The Venn diagram (figure 4.6) shows, that the predicted oscillating genes of the replicates overlap remarkably. The number of oscillating genes was reduced as more replicates are taken into account. An obvious assumption is that the replicates reduce noise and the impact of biological variation. So the genes which were consistently identified as rhythmically expressed seem to be the most reliable ones. Based on these results we decided not to use the gene expression profiles of the replicates but to combine randomly the expression values to new time-series. The procedure is described in the next chapter.

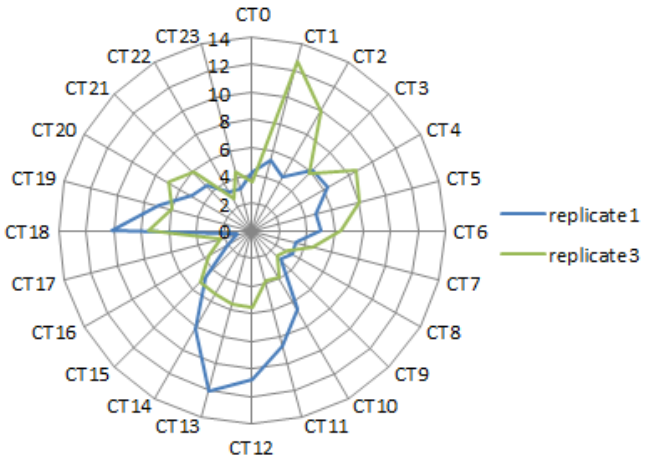

Figure 4.5 CT group distribution of replicate 1 and 3 under constant light condition.

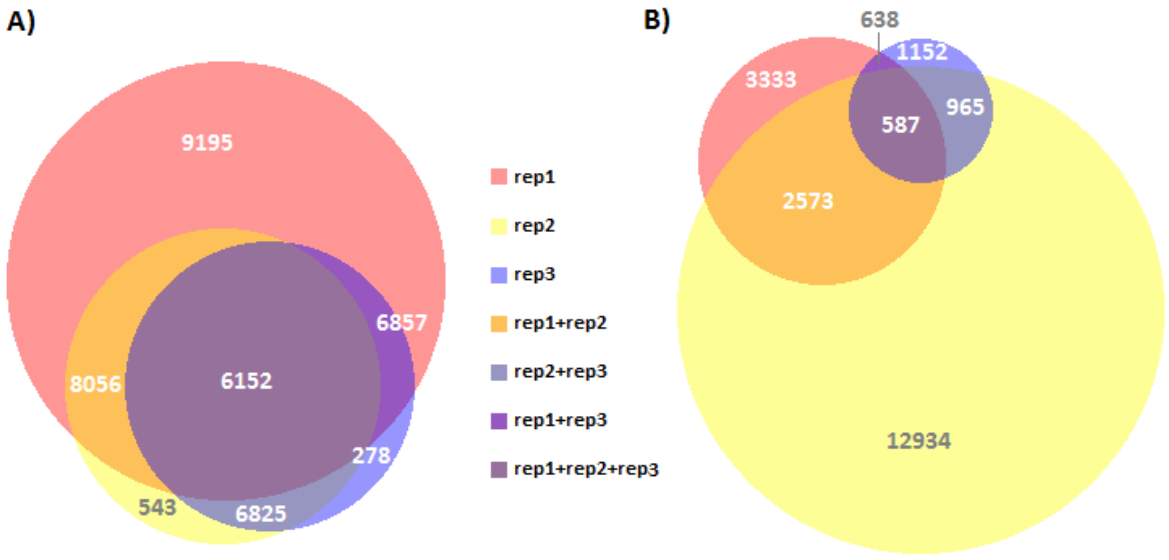

Figure 4.6 Area-proportional Venn diagrams address the absolute number of oscillating genes of three replicates and under two different light conditions

A) The microarray data of three biological replicates (LD condition) were analyzed by ARSER (FDR q<0.05). A total of 17956 genes were identified by ARSER for replicate 1 (rep1), while only 9272 and 7808 genes were found for replicate 2 (rep2) and 3 (rep3) respectively. B) The microarray data of three biological replicates (LL condition) were analyzed by ARSER (FDR q<0.05). A total of 12934 genes were identified by ARSER for replicate 2, while only 3333 and 1152 genes were found for replicate 1 and 3 respectively. Venn diagram was generated by BioVenn tool [35].

4.3 Randomization

The analysis of three biological replicates revealed that the expression values vary a lot and as a result the number of predicted oscillating genes as well as the ZT-/CT- group distribution differs significantly between the three replicates. Normally replicates are necessary to reduce noise and to control the impact of biological variation. Not least the costs limit the number of replicates but also the tissue consuming procedure of sampling. Despite the low number of only three or even a single replicate for the most experiments the complete expected value range should be covered. Considering such a small number of measuring points outliers have a big impact on the results and may generate a distribution far from the real pattern. For instance, if we consider the CT group distribution of the three biological replicates (Figure 4.4 and 4.5) as an example, we can see that the

patterns vary greatly and completely different conclusions can be drawn. If we are forced to analyze only one of the replicates the result would barely reflect the general pattern. It would be biased by to many individual characteristics. To avoid such a misleading result we decided not to analyze the biological replicates themselves but to combine randomly the expression values to new time-series.

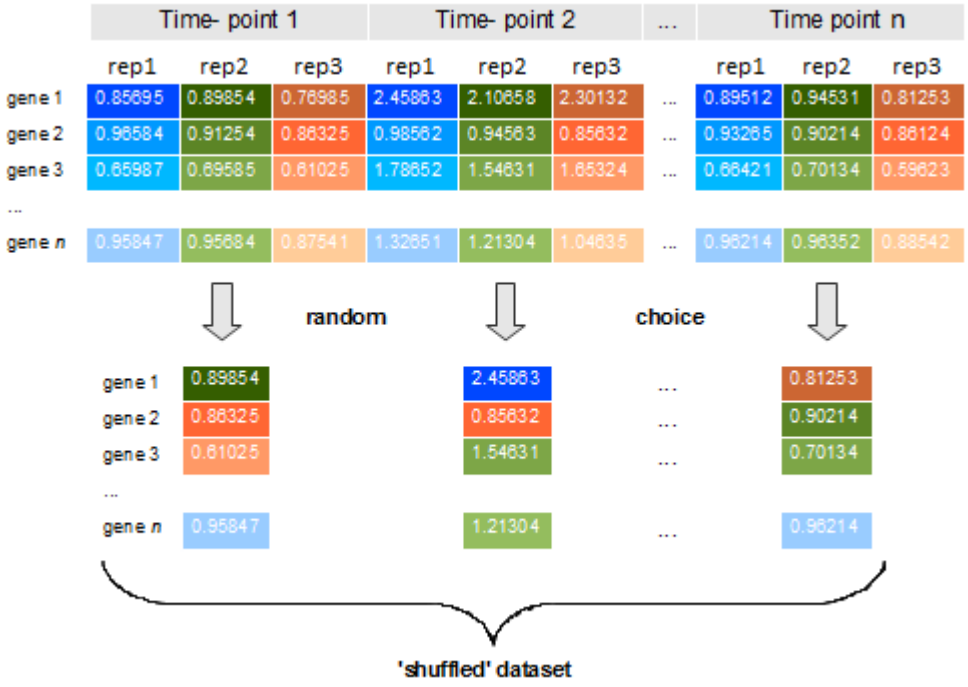

Figure 4.7 Schematic design of the randomization process
For every time-point and every gene we randomly selected one of the expression values of the three biological replicates. In the schema each replicate (rep1, rep2, rep3) has its own color to facilitate the assignment of the expression values from the new generated time series to the initial values. In principle for every gene an expression value is randomly selected in each time-point and these values were combined to form the new time-series.

The microarray time course data consists of 13 time- points, representing 48 h of observation obtained at 4 h sampling intervals for both light conditions (LD and LL). For every time-point and every gene we randomly selected one of the expression values of the three biological replicates. Figure 4.7 illustrates the procedure. In the schema each biological replicate has its own color to make it easier to follow the randomization step.

Replicate 1 (rep1) is highlighted in blue, replicate 2 (rep2) in green and the third replicate (rep3) in orange. Each expression value has an equal probability of selection and the time-points are treated independently of one another. Out of 61646 probes only 4 are shown as examples. For each gene an expression value was randomly selected from the three biological replicates per time-point and the values are combined to a new time series. The procedure was repeated 30 times for LD and LL condition data and we created 30 different time course series consisting of 13 time- points within a 4 h interval over a range of two days. The new time-series data was analyzed by ARSER with stringency threshold (q-value) set to 0.05. For each time-series the algorithm returned a list of predicted oscillating genes. When we compared the absolute number of oscillating genes it was notable that for all 30 time-series nearly the same number was predicted (Figure 4.8).

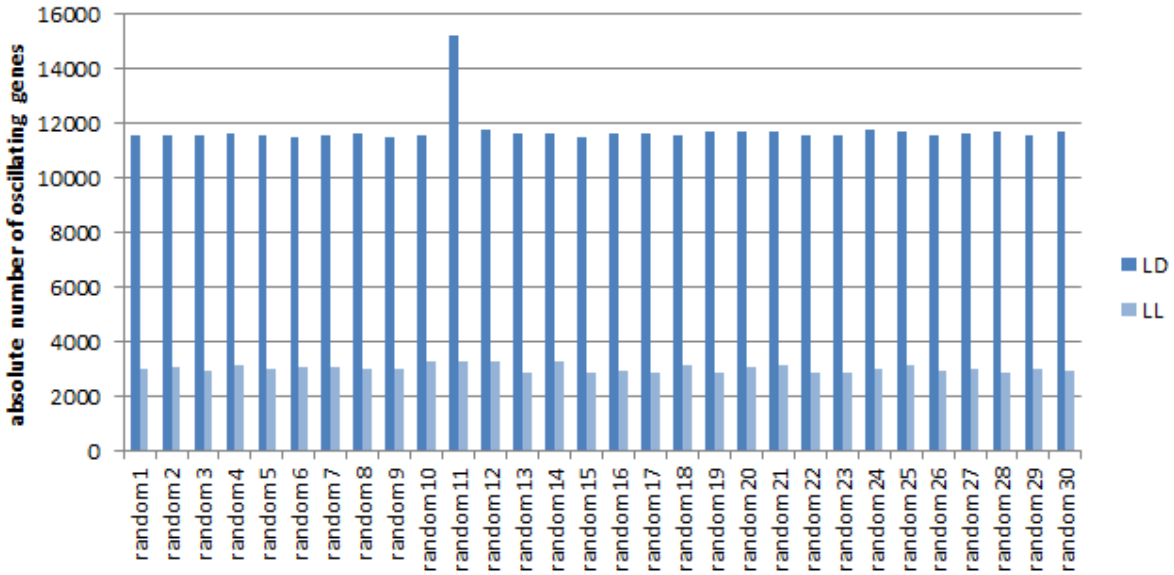

Figure 4.8 Absolute number of oscillating genes for LD and LL condition

The expression values of three biological replicates were used to generate 30 new time-series called "random1" to "random30" respectively. The time course series consist of 13 time- points within a 4 h interval over a range of two days. They were analyzed by ARSER (FDR q-value<0.05) and the number of oscillating genes determined.

Figure 4.3 shows the absolute and relative number of predicted oscillating genes for three biological replicates and they vary widely. The random combination of the replicates leads

to a more balanced pattern. We could not detect any outliers except the time- series called “random11” and we also kept this time-series in the dataset.

This result leads to the question whether all lists of oscillating genes contain the same genes. To test this we determined the number of genes identified in 30 out of 30 time-series. The overlap of all time-series was only 30.2 % (LD) and 4 % (LL) respectively. Based on Figure 4.8 this result was surprising, indicating that the lists contain a high number of false positives. In the next step we gradually reduced the stringency threshold and determined the number of genes identified if we compare the lists of 29, 28,…, 1 time-series (Figure 4.9).

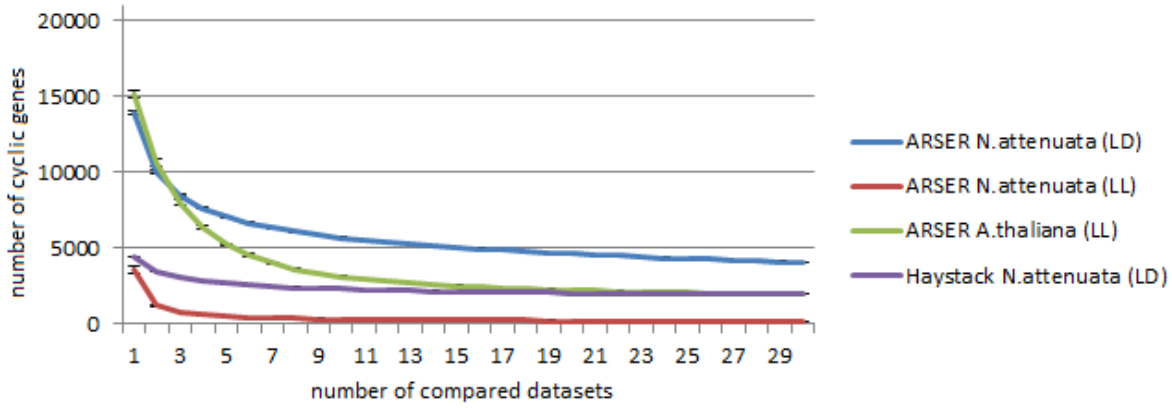

Figure 4.9 Absolut frequency of the average number of rhythmically expressed genes plot against decreasing number of compared time-series.

Two organisms, algorithms and conditions were tested. The blue, red and green curves visualize the results from the ARSER algorithm whereas the purple curve shows the result from the HAYSTACK algorithm. The genome of *N.attenuata* was analyzed by both algorithms and under light/dark as well as constant light conditions (blue, red, pruple curve). Arabidopsis was analyzed only with ARSER under LD conditions (green curve)

The more lists were used to calculate the overlap the less genes were identified as rhythmically expressed ones. We used different combinations of the lists to determine the average number of genes in the overlap. For instance, if the number of compared time-series is 5 then 5 lists out of 30 were randomly selected and checked to which extend the list entries overlap. This operation was repeated 30 times and the average of the overlap

calculated. So for each number of compared time-series 30 randomly chosen combinations have been considered. The average value for the overlap was entered in the diagram. At the beginning the curve falls sharply and approaches a remaining value after a number of about 20 compared datasets. 31.7% (LD condition) and 3.7% (LL condition) of the genes were identified as circadian in at least 20 of the 30 time series and we decided to choose these lists of common genes for further analysis, hereafter called LD_20 and LL_20. We first noticed the effect for our randomized time- series for *N. attenuata*. To demonstrate that the observed effect is common in microarray data analysis and not a result of the chosen algorithm we analyzed the randomized LD time-series with the HAYSTACK algorithm as well. As can be seen in Figure 4.9 the resulting (purple) graph shows the same characteristics. In order to prevent that it is an organism specific effect we generated 'shuffled' time course series for Arabidopsis on the basis of the experimental data of *Bläsing et al.* and analyzed the data also by ARSER. The resulting (green) curve shows a similar progression.

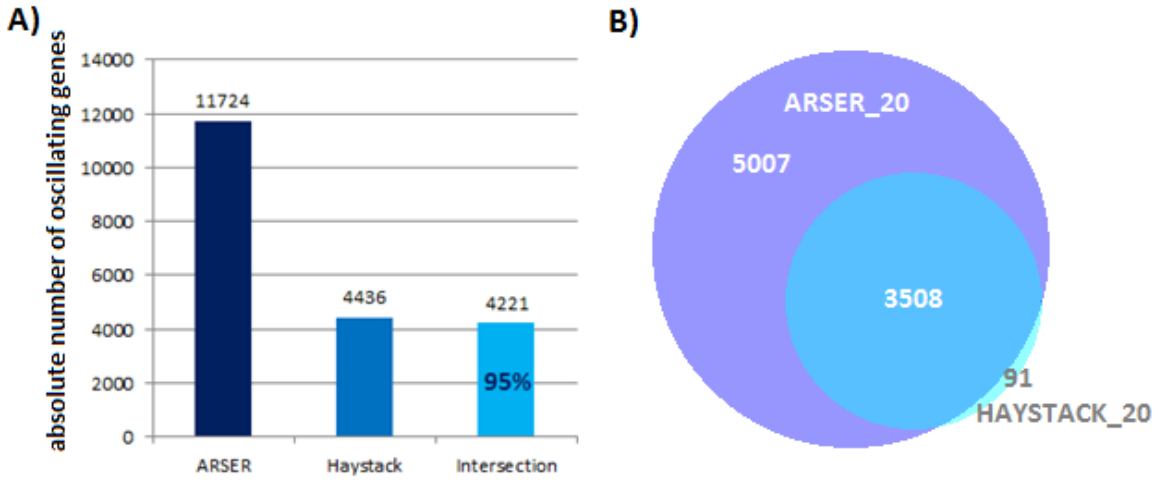

Figure 4.10 Compared performance of ARSER and HAYSTACK

A) Absolute frequency of the average values of genes classified as rhythmically expressed ones from ARSER, HAYSTACK and the overlap of both B) Area-proportional Venn diagram addresses the predictive power of ARSER and HAYSTACK algorithm. The detected genes had to show a periodic expression profile in at least 20 out of 30 investigated time-series. A total of 8515 genes were identified by ARSER and only 3599 genes were found by HAYSTACK. Venn diagram was generated by BioVenn tool [35]

Figure 4.9 also shows that ARSER predicts more rhythmically expressed genes. Therefore we decided to use this algorithm for further analysis. To validate the results we compared the performance from ARSER and HAYSTACK in more detail (Figure 4.10).

Both algorithms are based on the assumption that periodic patterns within expression profiles could not always be described with sine or cosine waves. Besides the fact that ARSER uses a laboriously procedure to identify the period, whereas Haystack assumes a predefined period of 24 hours. A further difference between both algorithms lies in the choice of their stringency thresholds. HAYSTACK considered a p-value<0.05 and ARSER a q-value<0.05.

Of all 26814 genes on the array ARSER classified on average 11724 genes as rhythmically expressed and HAYSTACK 4436. 95% of the cycling transcripts identified by HAYSTACK were also found by ARSER (Figure 4.10 A). In addition the gene lists ARSER_LD20 and HAYSTACK_LD20 were compared. ARSER_LD20 contained 8515 genes and HAYSTACK_LD20 3599. The overlap of both algorithms contained 3508 genes. So ARSER again detected 96% of the rhythmically expressed genes identified by HAYSTACK (Figure 4.10 B).

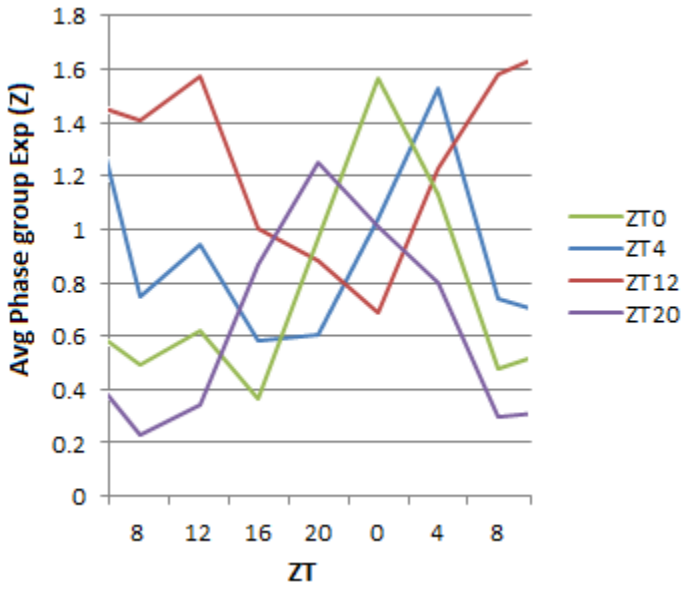

Figure 4.11 Phase group expression

The average Z scaled phase group expression of ZT0, ZT4, ZT12 and ZT20

5007 transcripts were uniquely identified as rhythmic by ARSER. We examined these genes to check whether their expression profiles showed really a periodic pattern and found that the majority shows a spike or rigid waveform. To give some representative examples Figure 4.11 shows the average Z scaled phase group expression of ZT0, ZT4, ZT12, and ZT20. The rhythm and peaking time within the

profiles were obvious. HAYSTACK seems to be more stringent, but Figure 4.11 shows that the algorithm may lose some important genes during the filtering steps. An adjustment of the cutoff values (correlation, fold, p-value and background cutoff) could lead to an optimization of the reported results, but such an analysis should not be part of this work.

4.4 Simulated data

The results of the randomized time-series leads to the hypothesis that genes which are identified as rhythmically expressed in at least 20 out of 30 time-series possess a high probability to be true positives. To test this hypothesis we generated simulated data with a known number of true and false positives, because for experimental data these values are unidentified. It should be proven that the arrangement to classify a gene as rhythmically expressed if it was identified in at least 20 out of 30 time-series reduces the number of expression profiles incorrectly classified as periodic. Each simulated dataset consist of 30 time series and 30000 genes with 12 time-points over a range of two days. The datasets mainly differ in the ratio of periodic and non-periodic profiles ranging from 5% to 15%. Stationary and non-stationary models were used to simulate periodic patterns as well as a combined model of sine and cosine functions. We used normally distributed white noise and AR processes of order one to simulate non-periodic time series. The simulated data were analyzed by ARSER and HAYSTACK to prevent algorithm specific effects. The resulting lists of genes classified as rhythmically expressed ones were compared and the overlap between an increasing number of lists calculated. For example the list with name "overlap_22" consists of a set of genes which are identified as oscillating in at least 22 out of 30 datasets. A confusion matrix, which contains the predicted and actual classifications (*Kohavi and Provost* [44]), was set up for each list of overlaps. The four fundamental members of the matrix are: true positives (expression profiles correctly classified as periodic), false negatives (expression profiles incorrectly classified as non-periodic), false positives (expression profiles incorrectly classified as periodic) and true negatives (expression profiles correctly classified as non-periodic). To evaluate the performance of

our randomization methodology we used different measures like accuracy, precision, sensitivity and specificity. These measures are defined by using the elements of the matrix and were calculated using the following terms

$$sensitivity = \frac{TP}{(TP+FN)} \quad (1)$$

$$specificity = \frac{TN}{(TN+FN)} \quad (2)$$

$$accuracy = \frac{(TP+TN)}{(TP+FP+TN+FN)} \quad (3)$$

$$precision = \frac{TP}{(TP+FP)} \quad (4)$$

where TP is the number of true positives; FN is the number of false negatives; TN is the number of true negatives; and FP is the number of false positives. In this case sensitivity, also called the true positive rate, provides information about the proportion of time series which were correctly classified as period patterns. For our simulated data the sensitivity decreases with an increasing number of compared lists (Figure 4.12).

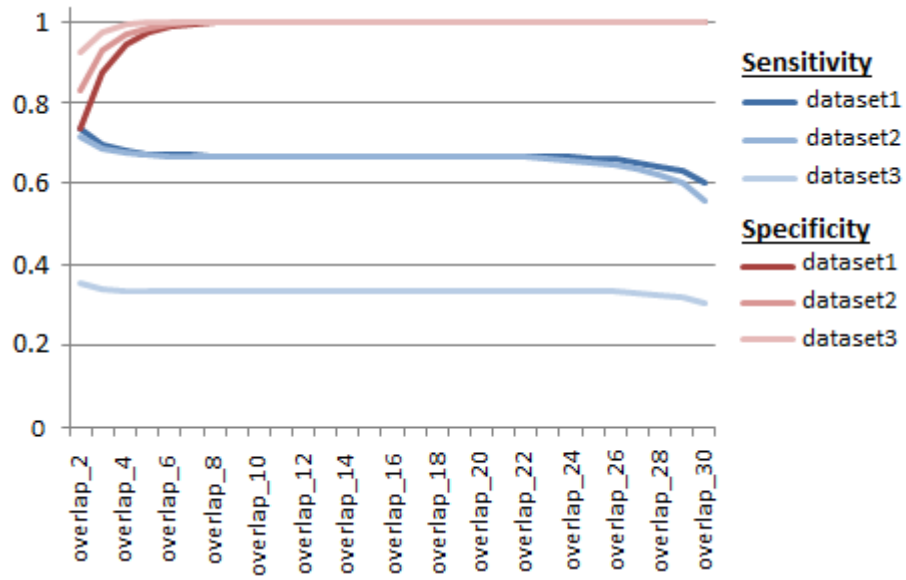

Figure 4.12 Performance evaluation measures for three datasets

Sensitivity and specificity for identifying periodic signals from three different datasets, whereas dataset1 consists of 4500 periodic and 25500 non-periodic, dataset2 of 3000 periodic and 27000 non-periodic and dataset3 of 1500 periodic and 28500 non-periodic expression profiles

This effect can be observed for all three datasets. The lower the percentage of periodic expression profiles, the smaller becomes the value. The smallest number of periodic signals can be found within the third dataset and a sensitivity of about 33% is reported. As important as the proportion of true positives is the proportion of true negatives measured by the true negative rate, called specificity (Figure 4.12). In this case a specificity of 1 means that no expression profile was incorrectly classified as periodic. For all compositions this value is reached when a gene is identified as rhythmically expressed in at least 8 of the 30 datasets. Here it is the quite opposite way compared to sensitivity. The specificity is higher for compositions with a smaller amount of periodic patterns. Based on this measures a gene could be classified as rhythmically expressed when it is identified in at least 8 datasets.

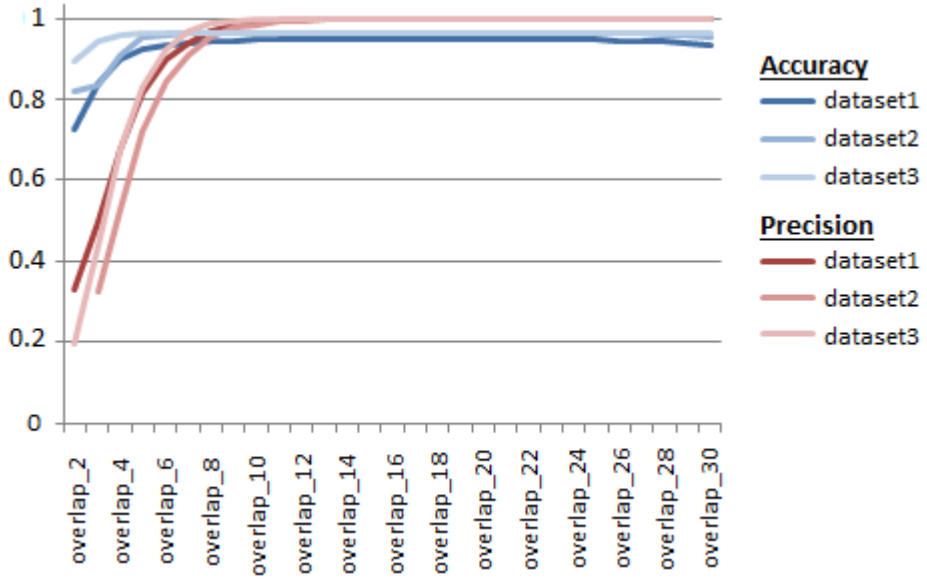

Figure 4.13 Performance evaluation measures for three compositions

Accuracy and precision for detecting rhythmic expression profiles within three different datasets, whereas dataset1 consists of 4500 periodic and 25500 non-periodic, dataset2 of 3000 periodic and 27000 non-periodic and dataset3 of 1500 periodic and 28500 non-periodic expression profiles

To confirm this assumption we also calculated the accuracy and precision. A high accuracy means that the number of genes correctly classified as periodic or non-periodic is close to the true value. Slightly different information is given by the precision value. This means

that more relevant than irrelevant results are returned. Both measures ascend with an increasing number of compared lists (Figure 4.13).

Especially the precision increases dramatically if the overlap is calculated between several lists, indicating that each added list reduces the number of irrelevant results. This assumption is verified by Figure 4.14. The number of false positives decreased dramatically as more gene sets are taken into account whereas the number of true positives stays nearly the same. If we take all measures of performance into account “overlap_20” reports the most reliable set of rhythmically expressed genes. The results were confirmed by HAYSTACK (data not shown). We did the same analysis using the HAYSTACK algorithm and the results for the performance evaluation measures were similar. Here, too, it could be shown that the number of false positives decreases as more lists were taken into account to calculate the overlap.

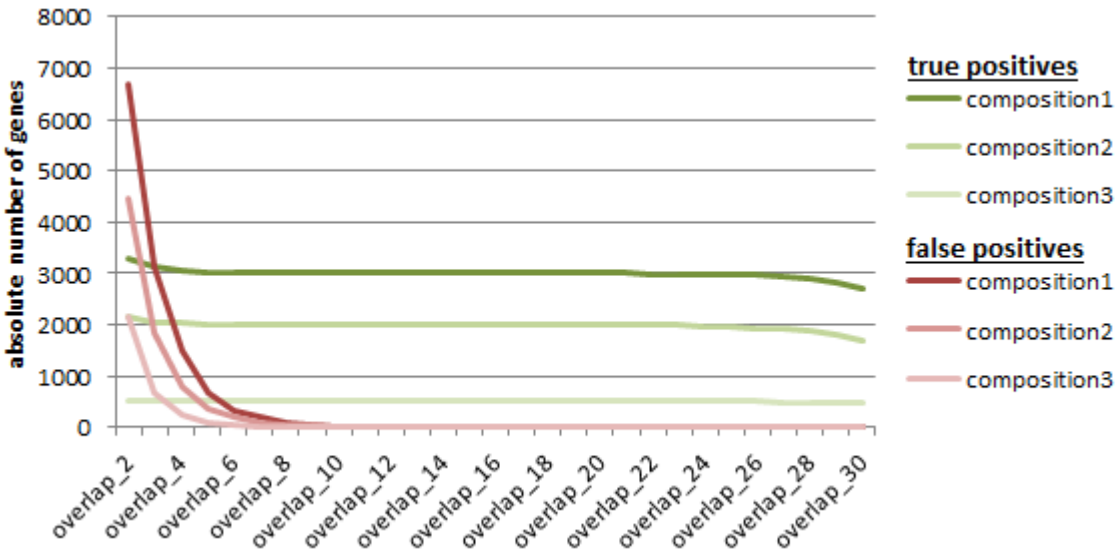

Figure 4.14 Absolute number of true and false positives for different ratios of periodic expression profiles within simulated time series

Number of true and false positives for different gene sets and amounts of periodic patterns, whereas composition1 consists of 4500 periodic and 25500 non-periodic, composition2 of 3000 periodic and 27000 non-periodic and composition3 of 1500 periodic and 28500 non-periodic expression profiles. The gene sets are named “overlap_x” which means that all these genes were identified as circadian by at least x out of 30 datasets.

We used the list of genes which were at least in 20 out of 30 datasets classified as rhythmically expressed to compare sensitivity, specificity, accuracy and precision between the average and randomization procedure (Figure 4.15). The average is often used when multiple replicates are available to infer the probable state of an average sample in the population. For each time-point the raw expression values of the replicates are averaged into a mean for every single gene. Following this procedure the values are normalized and logarithmized. To generate the randomized time-series we created 3 time-series which should perform as biological replicates as described in chapter3.3. Now the average was calculated of these time-series and analyzed by ARSER.

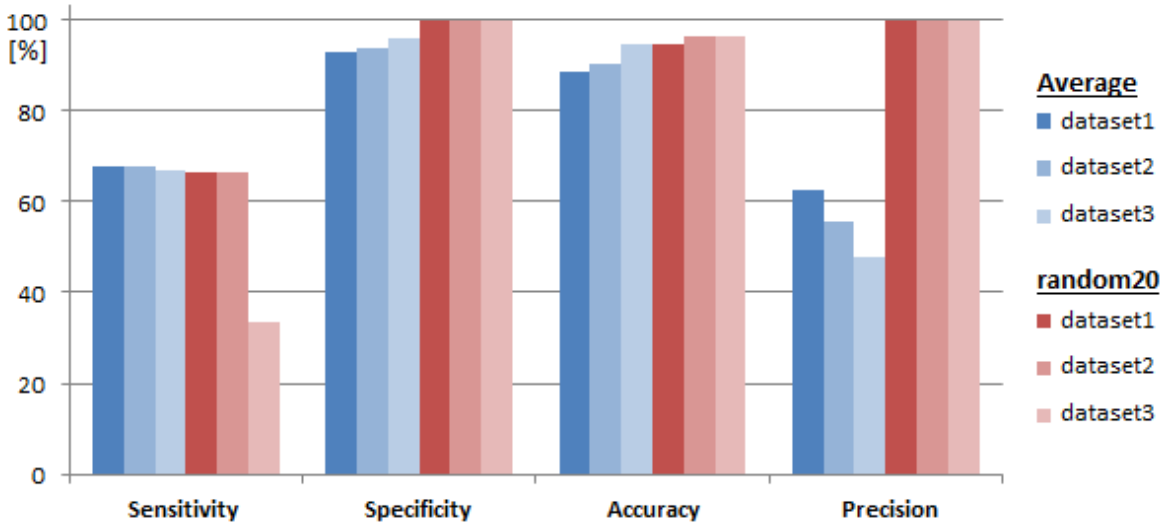

Figure 4.15 Performance evaluation measures for two data generation methods
Sensitivity, specificity, accuracy and precision for identifying periodic signals from three different datasets, whereas composition1 consists of 4500 periodic and 25500 non-periodic, composition2 of 3000 periodic and 27000 non-periodic and composition3 of 1500 periodic and 28500 non-periodic expression profiles.

The shuffled data lead to better results concerning specificity, accuracy and especially precision. Independent of the proportion of periodic signals the precision is always 100% for the random data. In contrast, the averaging procedure leads to decreasing precision as the ratio of non-periodic patterns is bigger. That means using the average expression value increases the probability to return irrelevant results. As mentioned above the sensitivity of the random data is very low for the third dataset, but for the first and second dataset the

randomization procedure outcompetes the average method in all performance measures. Although the average method returns always a higher number of true positives it also includes more false positives.

4.5 Different intervals for ZT and CT group determination

As mentioned in chapter 3.6 the proposed method from *Ueda et al.* [82] to determine the ZT and CT groups may lead to some distortion. We tested different intervals to analyze whether it is possible to avoid or reduce such falsifications. The analysis was done for the experimental microarray data of dark/light cycle as well as for the LL conditions. To take into account biological variation we used the three biological replicates for the analysis and determined the absolute frequency of genes per ZT/ CT group. We only considered genes which had the same ZT/ CT group in every replicate. At first we shifted the borders of the time bins and compared the results (Figure 4.16). *Ueda et al.* defined the borders of CT0 as (0, 1] and of CT1 as (1, 2] and so on. The same definitions were given for the ZT groups [82]. We used a shift of 0.5. ZT0 now contains genes with a peaking time in the interval (23.5, 0.5]. Genes with a peaking time between 0.5 and 1.5 were grouped to ZT1 and so on. The upper limit belonged to the currently regarded group and the lower limit to the previous group. For example a gene that showed maximal expression at 0.5 belonged to ZT0.

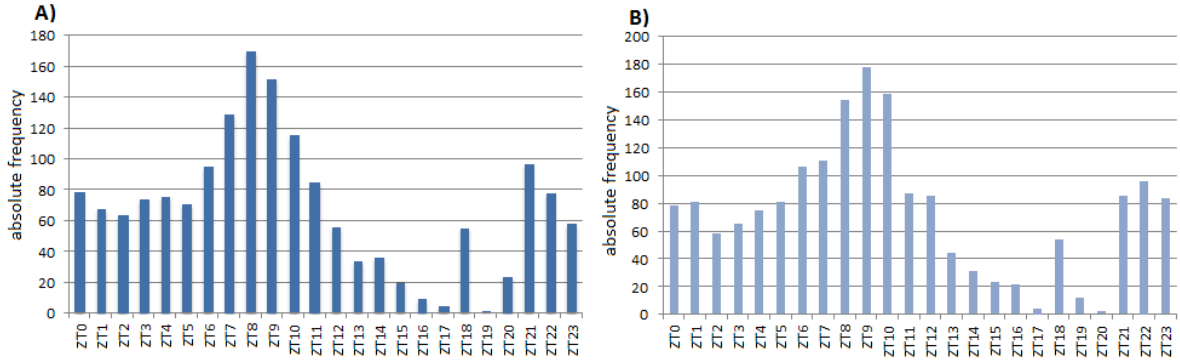

Figure 4.16 Absolute frequency of genes per ZT group under LD condition
A) Genes are assigned to ZT groups with an equidistant 1h interval B) Same interval of 1h but now the boarders are shifted

The Figure 4.16 indicates that the shift didn't change the ZT group distribution so much. The total number of genes per ZT group was a little bit higher for the interval with the shifted borders. The pattern of ZT group distribution stayed the same.

The next step was to duplicate the interval, because a reduction of the interval would be inappropriate. The new groups didn't met the definition of a ZT/CT group any longer and so we renamed the groups from "group0" (Gr0) to "group11" (Gr11). Gr0 contained genes with a peaking time between 0 and 2, Gr1 between 2 and 4 and so on. As expected more and more genes belonged to the same, new defined groups. The total number of genes per group was increased and the pattern of the distribution nearly stayed the same compared to the 1h interval (Figure 4.17 A). We tried to improve the disposition by yet another extension of the interval to an equidistant 4h interval (Figure 4.17 B). The 4h interval changed the ZT group distribution so that four out of six groups nearly contained the same number of genes. The previous observed pattern was not so obvious in this case. Although the extension of the interval increases the number of genes taken into account the borders are still fixed.

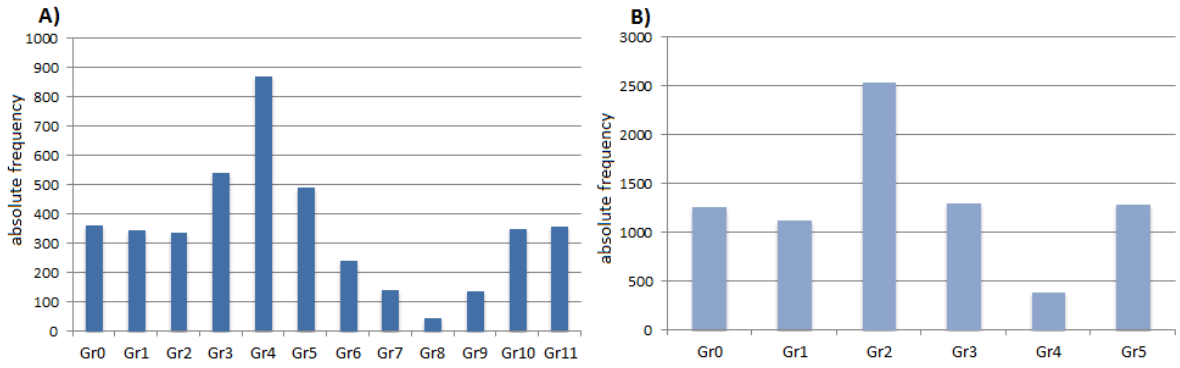

Figure 4.17 Absolute frequency of genes per group under LD condition

A) Genes are assigned to groups with an equidistant 2h interval B) Genes are assigned to groups with an elongated interval of 4h

Additionally the extension considered the neighborhood only in one direction. Therefore we decided that it would be better to determine the ZT/CT groups of the genes by using the 1 h interval.

4.5 Cluster analysis

The ARSER algorithm was developed to handle time course series with the common number of 12 to 13 time-points. We used the Arabidopsis dataset provided by the website *http://bioinfo.cau.edu.cn/BioClock/* to analyze the impact of a decreasing number of available measure points on the amount of genes classified as rhythmically expressed (Figure 4.18).

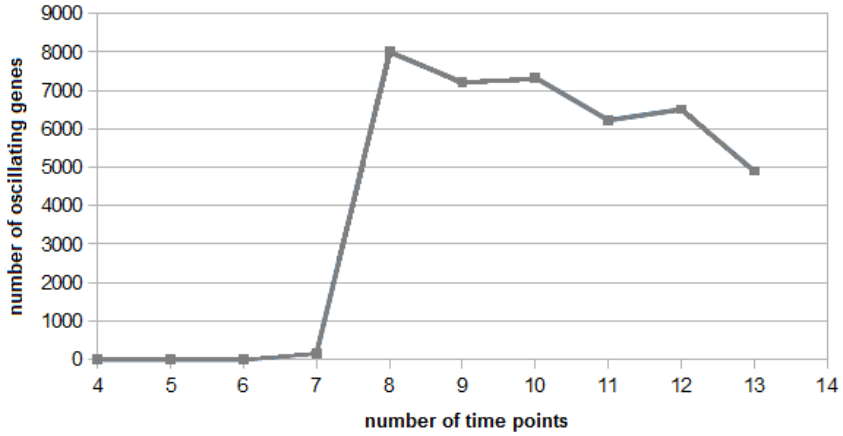

Figure 4.18 Absolute number of oscillating genes according to the number of time points in time series Time-series with an increasing number of time-points, starting with 4, were used as inputs for the ARSER algorithm and the absolute frequency of the average value of detected oscillating genes estimated.

As shown in the diagram above ARSER can be applied only to time course series with at least eight time-points. To investigate the expression of circadian regulated genes in different tissues the dataset from *Kim et al*. [42] was used. The time series within this dataset contained only six time-points thus it was not possible to analyze the data with ARSER. Therefore, the oscillating genes and especially the ZT groups had to be determined with another method. Based on the proposed principle from *Ernst et al.* [18] model profiles were used to apply clustering. The most important step was to design profiles for each ZT group as exactly as possible. It was assumed that a high correlation between the model profiles and the experimentally achieved time course series indicates the existence of a periodic pattern within the expression profiles automatically. Therefore, the better the model profiles the more accurately genes are classified as rhythmically expressed and

assigned to the right ZT group. Based on the ZT group determination for the artificially created datasets described in chapter 3.7 the median of phase, amplitude and period was estimated for every ZT group. These parameters were used to design specific model profiles for the time bins. We randomly selected a dataset from the LD condition to determine the parameter values and to design the model profiles. The profiles form the basis for the clustering process and were mapped to the expression values by calculating the correlation coefficient. To verify our profiles we calculated the correlation coefficient between the ZT group specific profiles and the median ZT phase group expression. Although individual genes within a ZT group displayed some variations each group had a single peak of expression per day. All genes within one group were very tight around the median across the time course. This demonstrates that analyses of the median expression patterns of the different ZT groups are sufficient to estimate the behavior of all genes within one group.

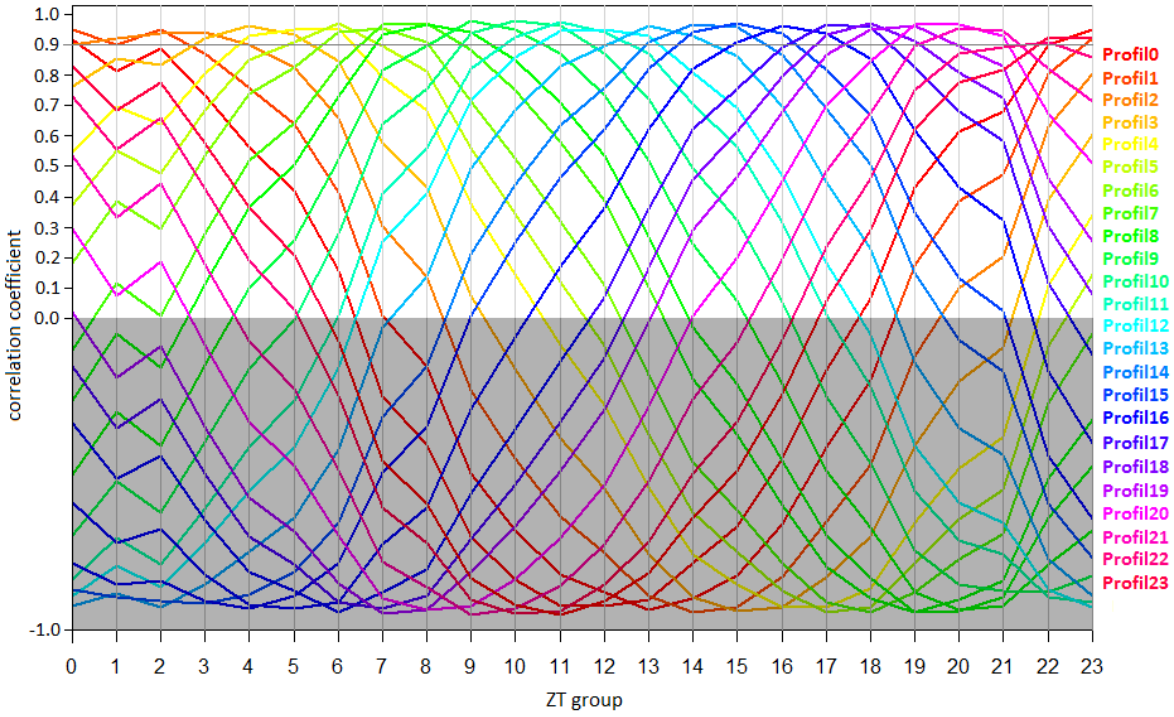

Figure 4.19 Performance evaluation measures for two data generation methods
Sensitivity, specificity, accuracy and precision for identifying periodic signals from three different datasets, whereas composition1 consists of 4500 periodic and 25500 non-periodic, composition2 of 3000 periodic and 27000 non-periodic and composition3 of 1500 periodic and 28500 non-periodic expression profiles.

So we mapped the model profile designed for ZT0 to the median expression profile of ZT0, ZT1, ..., ZT23 and calculated the correlation coefficient. This analysis was done for every model profile. In Figure 4.19 the correlation coefficients of the profiles are plotted against the ZT group. The resulting profile specific curves are illustrated in rainbow colors.
We were only interested in positive correlations for which reason the negative part of the curves is darkened in the diagram. Figure 4.19 shows that adjacent ZT groups had nearly the same correlation coefficient and therefore an exact determination of the ZT group seems to be impossible. The time bins could only be estimated with an accuracy of ±1. However, this precision of ±1 hour is remarkable, because the data was achieved from a 4h sampling interval. With common methods it is difficult to receive such a good resolution. Because of the limitation in accuracy the algorithm returned the ZT groups with the three highest correlation coefficients. Based on Figure 4.19 it is obvious that a threshold below 0.7 or above 0.9 for the correlation coefficient would be not successful. For this analysis we assumed high values for the correlation coefficient and therefore a threshold of 0.95 would also be appropriate.
We tested different thresholds for the other artificial datasets and the results are displayed in Figure 4.20. In Figure 4.20 A the total amount of genes which were mapped to one of the model profiles is shown in blue. As expected the number increases with a decreasing threshold for the correlation coefficient. Nevertheless the number of rhythmically expressed genes thus the amount of genes which was assigned to a model profile, is much smaller than the result of the analysis with ARSER. The ZT group determination was already done for every single dataset as mentioned earlier. This classification was used as a reference to test the accuracy of the new method. The absolute amount of genes which were mapped to the wrong ZT group is shown in red whereas genes with the right ZT group are illustrated in green. The green bars were analyzed in more detail in Figure 4.20 B. As mentioned above for every gene three ZT groups were reported and if one of these groups was the right one the classification was correct. Figure 4.20 B shows the exact amount of genes on the different positions. In more

than 90% of the cases the right ZT group was reported within the first two positions. If the correct ZT group was solely returned on the second position the correlation coefficient of the first and second place were very similar. Only in some exceptions the third position occupied the right ZT group. In these rare cases the correlation coefficients of all three places were very similar. A correlation coefficient of 0.9 was chosen for the classification of the dataset from *Kim et al.*

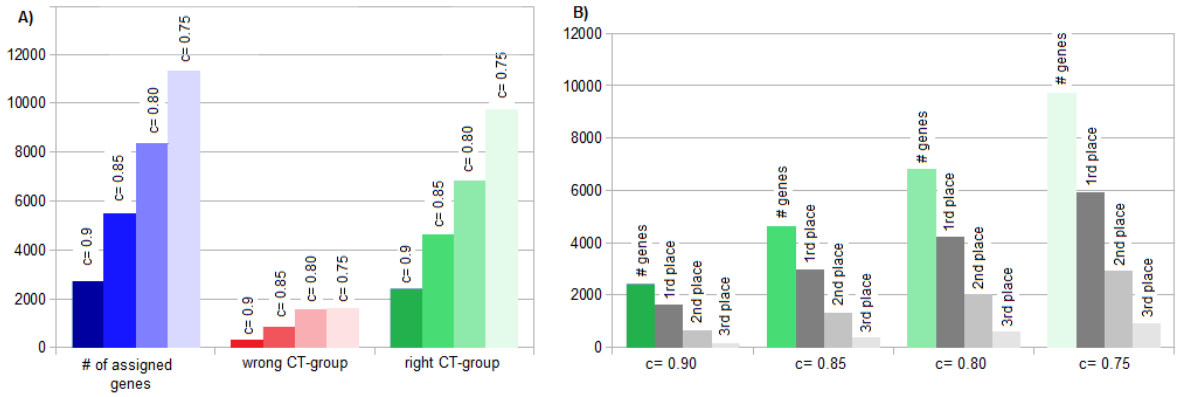

Figure 4.20 Performance of the clustering algorithm

A) The blue bars show the absolute frequency of assigned genes from the algorithm for different correlation coefficient thresholds. The absolute frequency of genes which were assigned to the wrong ZT group with various correlation coefficients as threshold is shown in red. Green indicates the absolute frequency of genes assigned to the correct ZT group also for different thresholds. B) The algorithm always returns the three ZT groups with the highest correlation coefficient for every gene. The green bars indicate the absolute frequency of genes assigned to the correct ZT group with different thresholds. Gray shows the absolute frequency of genes placed at the various positions as part of the whole assigned number. The ZT group at first place shows the highest correlation coefficient.

4.7 Comparison LD 12:12 and LD 16:8

Plants benefit greatly from a circadian clock that adjusts their overall metabolism in anticipation of the highly predictable arrival of sunrise and sunset [37]. By exact adaption on prevailing conditions an advantage concerning the plants fitness arises and these organisms automatically have a selective advantage [16]. Thus, to achieve an optimum performance the activation of genes which are involved in light dependent pathways has to be accurately timed. A desynchronization of the natural day/night cycle

results in higher costs for the plants. Based on the results from *Daan et al.* [16] for mammalian the model of two separate circadian oscillators (morning and evening oscillator) that drive activity was taken up. The "morning oscillator" (M) is accelerated by light and synchronized to dawn, whereas the "evening oscillator" (E) is decelerated by light and synchronized to dusk [30]. *Todd et al.* [55] were able to show that this model also works for plants. Most of the genes were expressed at a specific time of day and this time-point could be influenced by external stimuli. Most of the organisms show two activity bouts during the day, one in the morning and one in the evening [1]. In Arabidopsis most of the circadian genes peak right before dusk and dawn respectively. *Todd et al.* [56] compared the gene expression of plants grown under long-day as well as short-day conditions. Genes from the long-day plants reached their maximal expression later and in doing so, they were responding to the changed light conditions and elongation of day-length. To test whether this is also true for *N.attenuata* we compared the gene expression of plants grown under light- dark cycles with 12 h light and 16 h light respectively. The time course series from *Kim et al*. [42] were used to analyze genes under elongated day-length condition. ARSER cannot be used to analyze time-series with only 6 time-points therefore we edited the data. To receive time- series with 12 time-points in a 4 h interval we simply duplicated the data and concatenated the time-series to a two day time course. In the next step, the randomization procedure (chapter 4.3) was applied to these time-series to generate 30 new time-series. As a reference set we used the LD_20 list introduced in chapter 4.2. LD_20 contains genes identified in at least 20 out of 30 randomized time-series of plants grown under 12 h light and 12 h dark condition. The gene names of the dataset from *Kim et al.* and the LD_20 list were compared and we collected those which occur in both. The concatenated time-series of these genes from the Kim dataset were analyzed by ARSER to check whether these genes were also rhythmically expressed in the different tissues. In the next step the ZT groups were determined by calculating the molecular peaking time as described in chapter 3.5 as well as by applying cluster analysis described in the former chapter. The clustering worked with

only 6 time-points. Figure 4.21 illustrates the results of the comparison between molecular peaking time method and cluster analysis to determine ZT groups.

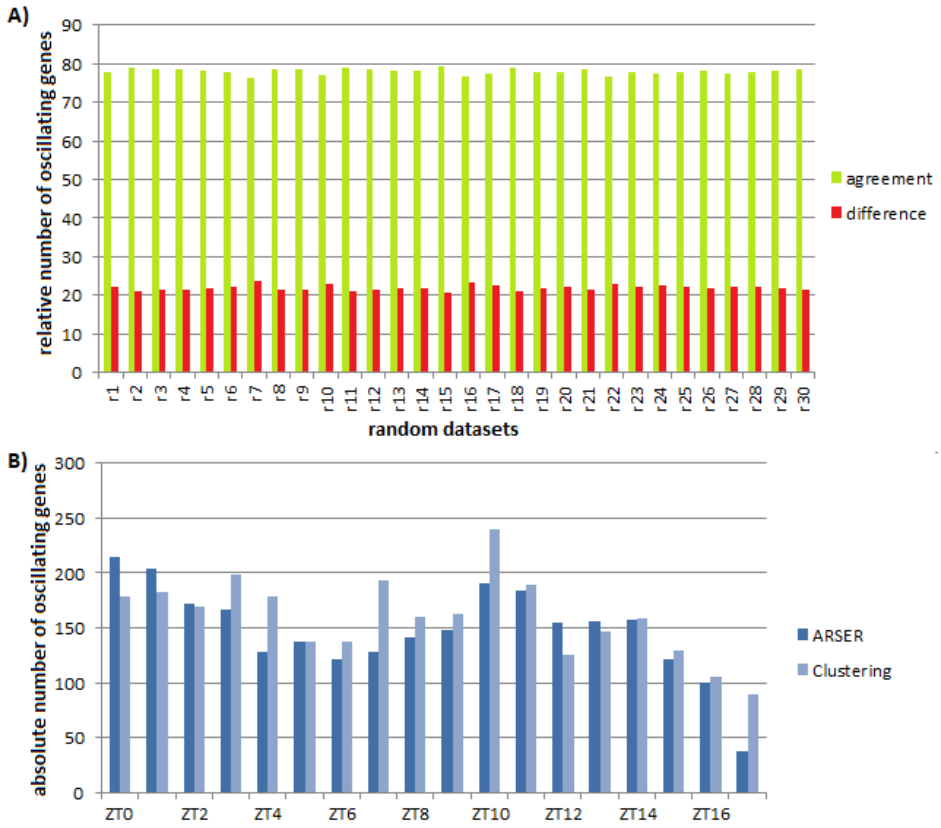

Figure 4.21 Comparison of molecular peaking time method and cluster analysis to determine ZT groups A) relative frequency of genes which received the same ZT group from both methods B) ZT group distribution achieved by both methods for the genes in common

However the absolute amount of rhythmically expressed genes depends on the method. The ARSER algorithm detects on average twice as much genes as the clustering algorithm. Based on this result we decided to use the gene lists returned from ARSER.

The gene expression profile varies between the individual time-series, leading to a different ZT group determination for one gene. To compare the ZT groups of both light conditions it was necessary to assign every gene a single ZT group. To achieve such a classification across all datasets the mode (the number that is repeated more often than

any other) was used. We decided to use this parameter, because it is easy to determine and less influenced by outliner. Figure 4.22C shows the resulting ZT group distribution for both light conditions. It is notable that most of the genes under LD 12:12 condition are assigned to ZT groups 3, 5, 8, 18 and 21. The elongation of day-length seems to shift the distribution slightly and now groups 11, 14, 18 and 23 contained the highest number of genes.

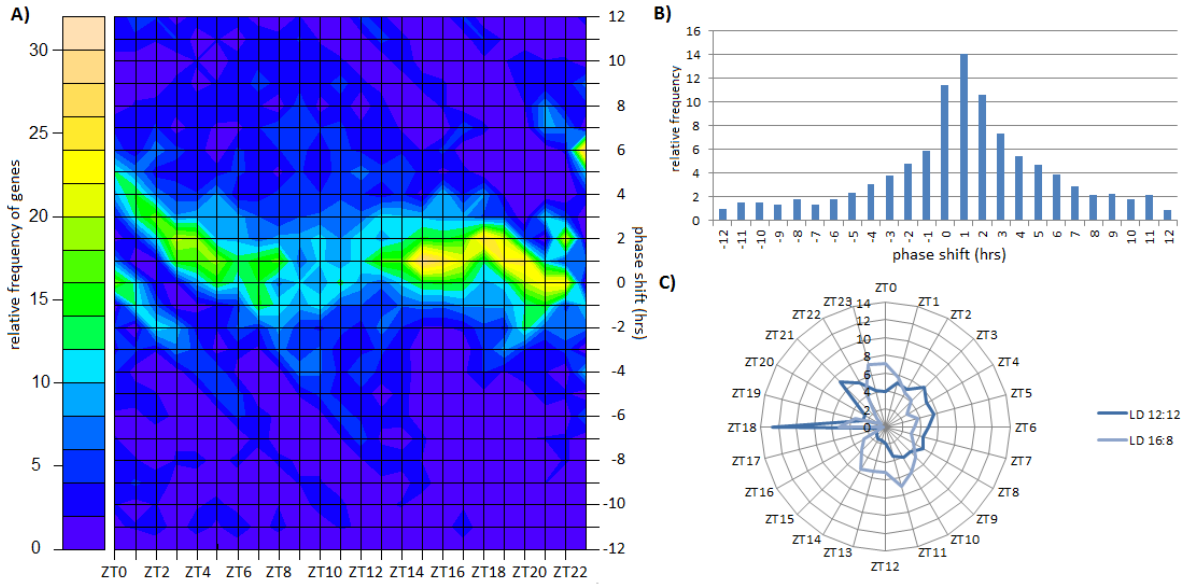

Figure 4.22 Comparison of the gene expression in longday and 12h light/12h dark cycles

A) Time shift topology graph plots percent of genes ZT group shifted per time bin (y-axis) by the reference ZT group (x-axis). Percent of genes was calculated as the number of genes with a given time shift per ZT group divided by the total number of genes within that ZT group. A positive ZT group shift reflects a later ZT group than the reference condition and a negative one reflects an earlier ZT group than the reference condition. LD 12:12 was taken as reference. B) Relative values of rhythmically expressed genes with the same time shift compared different conditions. ZT group classification under LD 12:12 condition was taken as a reference set. Difference was calculated by substituting the longday values from LD 12:12 values within an interval from -12 to 12 in which ZT0 was set as 0 C) Relative number of genes per ZT interval for both light conditions

On the whole, the amount of genes per ZT group is much more balanced under LD 16:8 condition and also the variance is lower. In the next step we analyzed the behavior of the genes within ZT groups. To compare both conditions the ZT group determination under LD

12:12 condition was used as a reference set and the difference between LD 12:12 and LD 16:8 was calculated. The difference was estimated within an interval from -12 to 12 in which ZT0 was defined as central point. Figure 4.22B shows that most of the genes peaked with a delay of 1 or 2 hours under long-day condition. A time shift of 5 or more hours in either direction was hard to find. Interestingly, about 11% of the genes kept their ZT group. The surface diagram represented in Figure 4.22A visualizes the three-dimensional data in a 2D format. The colors correspond to the relative frequency of genes ZT group shifted per time bin (y-axis) by the reference ZT group (x-axis). As can be seen from the diagram two hotspots have emerged within the ZT groups. The first one relates in principle to the morning genes, while the second one spans the ZT groups of the evening genes. Most of the time shifts occurred within the ZT groups 15 to 20 correlating with the shifted time of dusk. Therefore, the evening genes under long-day condition reached their maximal expression 1 or 2 hours later. That is what we were expecting. The plants react to the elongated day length with a delayed activation of the evening genes. In contrast the expression of the morning genes is slightly advanced in anticipation of dawn. Such an adaption enables the plant to use the full capacity of day-light and as a consequence the plants fitness arises. However, a large part of the morning genes showed a delayed expression. This might be due to the fact that oscillators are coupled. Thus, the gene expression is synchronized by internal signals and the rhythm of these genes should persist under LL condition. To test this hypothesis, we generated an area-proportional Venn diagram (Figure 4.23). The morning gene group contained genes of

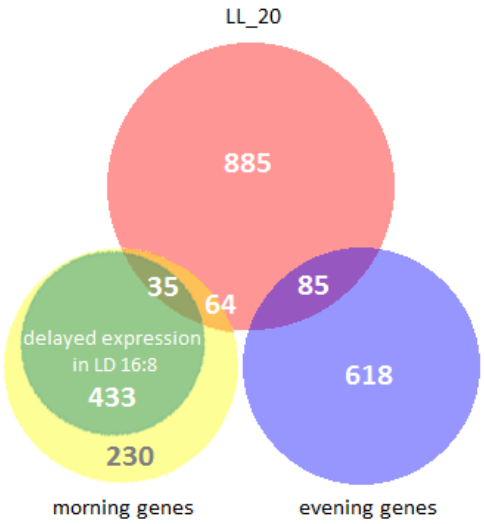

Figure 4.23 Area-proportional Venn diagram addresses the morning and evening genes and the genes of the LL_20 list. A total of 1034 genes were listed in LL_20. 703 morning and 725 evening genes could be identified. Within the morning genes 433 showed a delayed expression. Venn diagram was generated by BioVenn tool [35]

ZT22, ZT23, ZT0, ZT1 and ZT2. For the evening group we selected the genes with ZT10 to ZT14. The diagram shows that only a small part of morning and evening genes kept their diurnal rhythm in LL. Likewise, only a small number of the delayed morning genes could be found in LL_20.

4.8 Comparison leaf LD 16:8 and root LD 16:8

A number of studies had shown that plants have tissue specific clocks [79, 80]. This raises the question how the various clocks are synchronized. One of the first answers which come in mind is that the rhythm has to be entrained. In principle there are two options: entrainment via internal or external stimuli. Central circadian pacemakers and communicated rhythmic signals have little influence on plant rhythms [80]. This observation leads to the conclusion that environmental cues regulate the variety of internal clocks. The most promising candidates for that task are the Zeitgeber light and temperature [28]. As aforementioned, *Kim et al.* [42] used microarray experiments to analyze different tissues of *N. attenuata*. The resulting dataset contained gene expression data from the root and leaf tissue. In the year 2008 *James et al.* [36] proposed the regulation of the clocks in the root tissue by a photosynthesis-related signal from the shoot. Thus, the plant clock is organ-specific but not organ-autonomous. They analyzed different well-known components of the internal clock in root and shoot tissue of *A. thaliana*. At first they used the COSOPT algorithm to detect rhythmically expressed genes in both tissues. 13.7% of the genes showed a periodic expression pattern within the shoot whereas only 3.2% of the genes cycle in the root. We repeated this analysis for *N. attenuata* with the dataset of

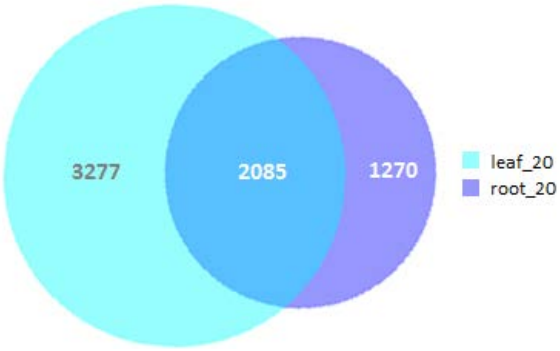

Figure 4.24 Area-proportional Venn diagram addresses the absolute number of rhythmically expressed genes for leaf and root tissue. Venn diagram was generated by BioVenn tool [35]

Kim et al. and could confirm the results from *James et al*. ARSER cannot be used to analyze time-series with only 6 time-points (Kim dataset) therefore we edited the data. To receive time- series with 12 time-points in a 4 h interval we simply duplicated the data and concatenated the time-series to a two day time course. In the next step, the randomization procedure (chapter 4.3) was applied to these time-series to generate 30 new time-series. Afterwards we selected the genes which occur at least in 20 of the 30 time series (leaf_20, root_20). Leaf_20 contained 5362 genes whereas root_20 encompasses 3355 genes. The determined percentages significantly exceeded the values of *James et al.*, but they investigated plants in LL condition. In the root tissue the relative frequency of rhythmically expressed genes was 12.5% and 20% in leaf tissue (Figure 4.24). Also the ratio between leaf and root tissue differs with ½ for *N. attenuata* in contrast to ¼ for Arabidopsis. A total of 2085 rhythmically expressed genes overlapped between the two tissues. Therefore, more than 50% of the genes rhythmically expressed in the root tissue show a recurring rhythm in leafs. It is also remarkably that the oscillating gene expression of some genes seems to be organ specific. An amount of 3277 genes were rhythmically expressed only in leafs and 1270 only in roots. Afterwards ZT groups were determined according to the maximal gene expression of each gene within a day as described in chapter 3.6 and 3.7. It was analyzed how this classification differs between both tissues. Therefore the difference between the ZT groups for a gene in both tissues was calculated whereas the root tissue was taken as reference. To visualize the discrepancy, the difference was calculated within an interval of -12 to 12 as described in chapter 4.6 and the relative frequencies of the time shifts are plotted in a histogram (over all ZT groups, Figure 4.25 B) and surface diagram (within single ZT groups, Figure 4.25 A). The distribution of the CT groups was similar in both tissues (Figure 4.25 C). Only the absolute amount of genes per ZT group differed. In the light of this diagram, it seems that only small changes in the ZT group classification between both tissues occur. Figure 4.25 B underlined this impression. Most of the genes show the same peaking time or at least a shift of 1 or 2 hours, but all possible shifts within the 1h interval were represented. The

surface diagram of Figure 4.25 A shows the time shift for every single ZT group. The detected pattern confirmed the assumption that only small changes occur. Genes within ZT groups ZT0- ZT7 and ZT12-19 showed predominantly a negative time shift which means that the genes peaked earlier in in the leaf tissue than in the root. ZT8-ZT11 showed no real trend in one direction. The groups contained genes which peaked earlier in the leaf as well as genes which peaked earlier in the root.

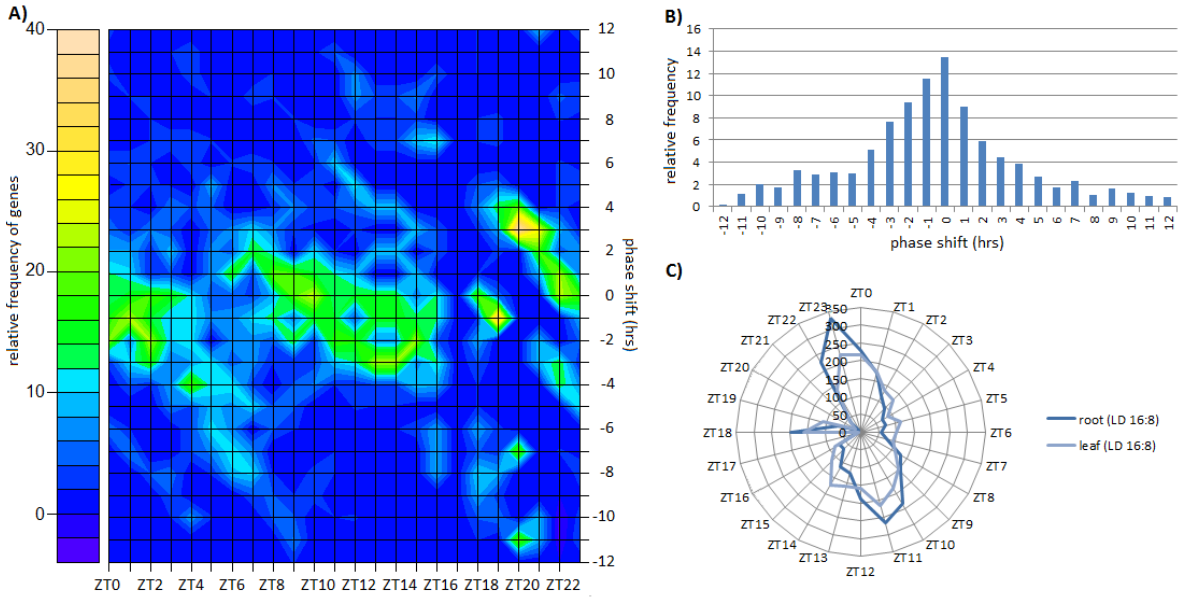

Figure 4.25 Comparison of the gene expression in leaf and root tissue

A) Time shift topology graph plots percent of genes ZT group shifted per time bin (y-axis) by the reference ZT group (x-axis). Percent of genes was calculated as the number of genes with a given time shift per ZT group divided by the total number of genes within that ZT group. A positive ZT group shift reflects a later ZT group than the reference tissue and a negative one reflects an earlier ZT group than the reference tissue. Root tissue was taken as reference. B) Relative values of rhythmically expressed genes with the same time shift compared different tissues. ZT group classification of root tissue was taken as a reference set. Difference was calculated by substituting the leaf values from root values within an interval from -12 to 12 in which ZT0 was set as 0 C) Absolute number of genes per ZT interval for both tissues

4.9 Comparison LD and LL conditions

To determine the number of rhythmically expressed genes in *N. attenuata* the expression profiles under LD and LL conditions were compared. As aforementioned, a

gene is classified as rhythmically expressed if the detected rhythm persists under constant light conditions. The microarray time course data consists of 13 time- points, representing 48 h of observation obtained at 4 h sampling intervals for both light conditions (LD and LL). It could be shown that the three biological replicates were very different and therefore we decided not to analyze the biological replicates themselves but to combine randomly the expression values to new time-series. For every time-point and every gene we randomly selected one of the expression values of the three biological replicates and combined them to a new time series. The procedure was repeated 30 times for LD and LL condition data and we created 30 different time course series consisting of 13 time- points within a 4 h interval over a range of two days. The time course series were analyzed by ARSER to identify rhythmically expressed genes. Figure 4.26A shows comparative the absolute and relative frequency of rhythmically expressed genes in LD and LL condition. The specified values relate to the average value of the 30 random time-series.

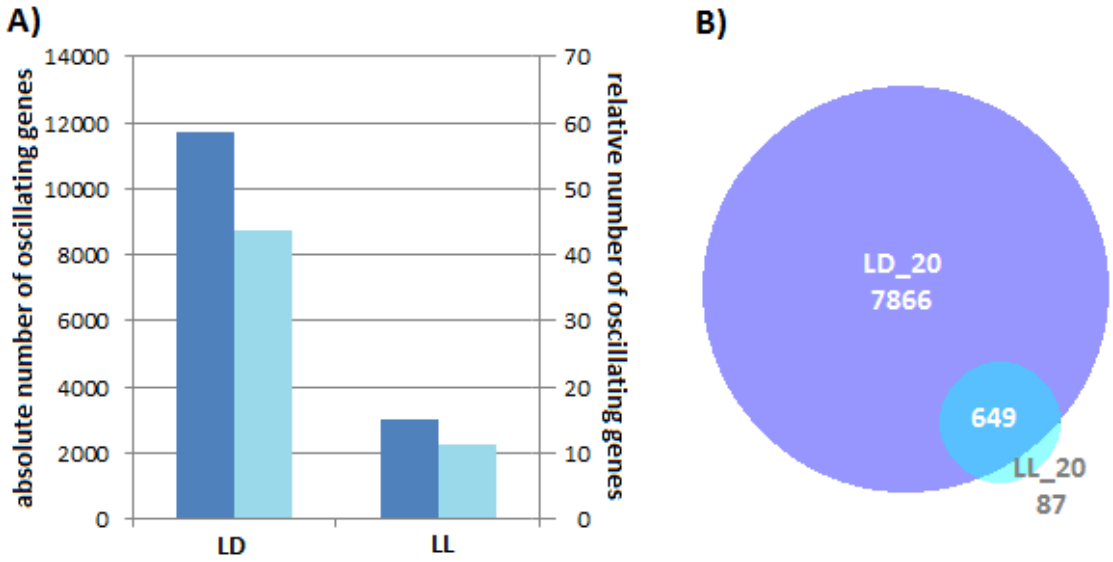

Figure 4.26 Number of rhythmically expressed genes under different light conditions

A) Absolute (dark-blue) and relative (light-blue) frequency of rhythmically expressed genes under LD and LL condition B) Area-proportional Venn diagram addresses the absolute number of oscillating genes under LD and LL condition. Here we used the lists of genes detected by ARSER as rhythmically expressed in at least 20 out of 30 time-series. A total of 8515 genes were identified under LD condition, while only 736 genes were found under constant light condition. Venn diagram was generated by BioVenn tool [35].

Additionally the overlap of genes between these two conditions was calculated and is depicted in Figure 4.26 B. Here we used the list of genes previously introduced as LD_20 and LL_20. As one can see the number of cyclic genes under LD condition is much higher than the identified number under constant light. More than half of the genes showed a diurnal rhythm. Based on the assumption that circadian rhythms persist in their 24-hour periodicity the detected genes under LL conditions have to be the ones with the most robust rhythm. On average 13.4% of the *N. attenuata* transcriptome seems to be regulated by the circadian clock. In comparison, Arabidopsis microarray time course data predicted a number between 6% and 15% [56]. The percentage rate fluctuates depending on the applied method and defined limitations. Thus, the LL_20 list leads to a number of only 3%.

Already the Swiss botanist de Candolle noticed that the free-running period of leaf opening and closing was shortened under constant light conditions by approximately 2 hours. The internal rhythm thus differs significantly from the 24-hour period of the Earth's light-dark cycles. He proved that entrainment by environmental cues was not the reason for the shortened period, because longer cultivation without external stimuli leads to a full desynchronization of the internal rhythm from the exogenous light/dark rhythm. Candolle's findings could be repeated later on by experiments which confirmed the existence of oscillations in the absence of environmental cues. The period of genes which were rhythmically expressed under LD and LL condition was examined for *N. attenuata*. Figure 4.27 shows the absolute (A) and relative (B) frequency of period shift in hours. The shift was calculated using the determined period under LD condition as a reference. A difference of -1 means that the detected period under LL condition is shortened by 1 hour compared to the period under LD condition. In about half of the cases the period length under LL condition is shortened compared to LD condition. Around 30% of the genes didn't change their period at all whereas only 20% of the genes showed an elongated period under LL condition.

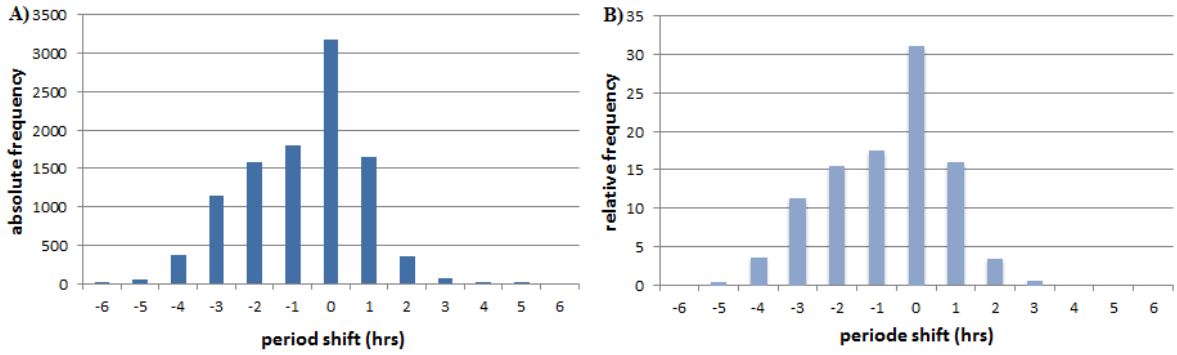

Figure 4.27 Comparison of the period of rhythmically expressed genes under different light conditions. The shift was calculated using the period under LD condition as a reference.
A) absolute frequency B) relative frequency.

In the next step we turned our attention to the amplitude of the genes which were rhythmically expressed under both conditions. As many of the selected genes had small amplitudes the comparison was made only taking into account genes with high amplitudes. The results are shown in Figure 4.28 and as one can see the amplitude decreased under constant light condition. The smaller the amplitude under LD condition the smaller was the difference compared to LL condition.

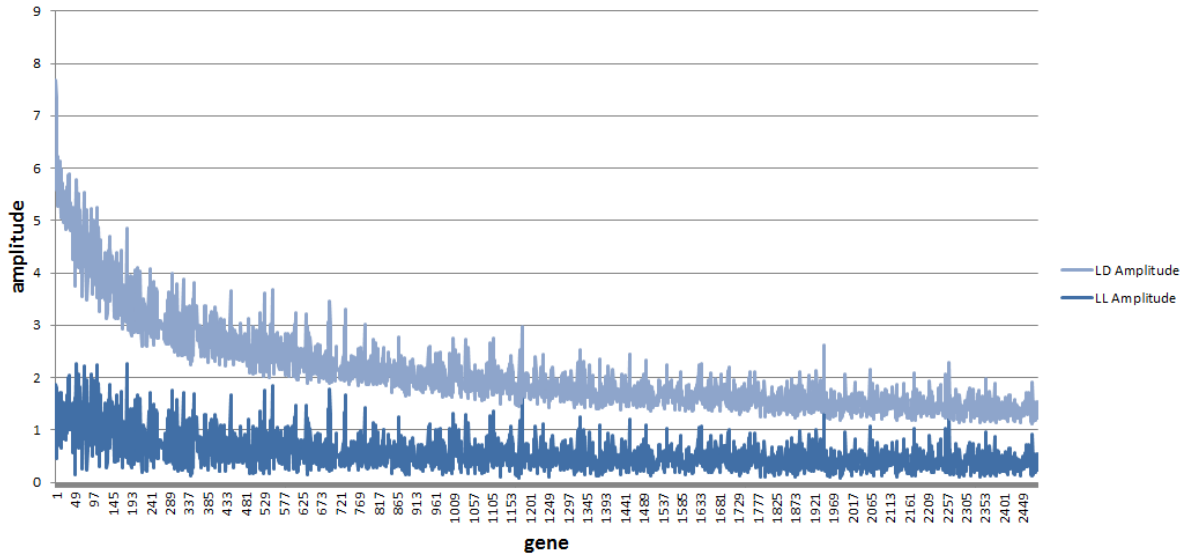

Figure 4.28 Amplitude of rhythmically expressed genes under different light conditions

Higher plants are multioscillatory systems, which are synchronized by environmental time cues [80]. This phenomenon is known as “internal synchronization” and the fact that the rhythms persist in the absence of external cues indicates the existence of a self-sustained endogenous oscillator [57, 52]. It was demonstrated, that photo cycles are an effective zeitgeber and are able to resynchronize rhythms [57]. If the cycles are shifted a time lag occurs until there is a new synchronization [23]. In constant light conditions there are no external cues; therefore, a new synchronization is not possible.

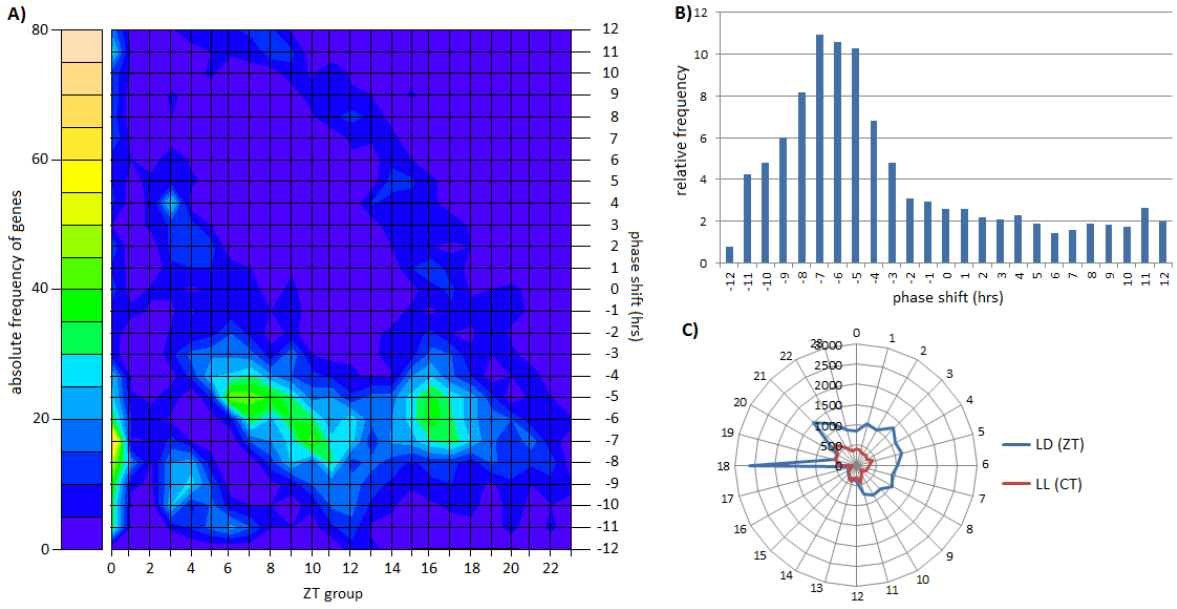

Figure 4.29 Comparison of the gene expression under LD (12 h light/12 h dark) and LL conditions
A) Time shift topology graph plots absolute number of genes CT group shifted per time bin (y-axis) by the reference ZT group (x-axis). A positive CT group shift reflects a later CT group than the reference condition and a negative one reflects an earlier CT group than the reference condition. LD condition with 12 h light/12 h dark was taken as reference. B) Relative values of rhythmically expressed genes with the same time shift compared in different conditions. ZT group classification of LD (12:12) was taken as a reference set. Difference was calculated by substituting the constant light values from LD (12:12) condition values within an interval from -12 to 12 in which CT0 was set as 0 C) Absolute number of genes per ZT/CT interval under LD (12:12) compared to LL

Figure 4.27 and 4.28 show that two important parameters (period, amplitude) had changed following the transfer. Now the interesting question arises whether such an

alteration had also an influence of the peaking time. To answer this question we compared the ZT and CT groups of the genes. At first we determined the ZT group for every rhythmically expressed gene under LD condition in all 30 time-series. Depending on the time course one gene was assigned to different ZT groups sometimes. To solve this problem we calculated the mode of the ZT groups for every single gene. These results were used as a reference set. The same steps were done for the genes under constant light conditions. Figure 4.29 C shows the ZT and CT group distribution for both light conditions. The distribution seems to be similar although the absolute number of genes per time bin was higher under LD condition. Finally we compared the ZT and CT groups of each gene under both conditions and calculated the difference in hours (Figure 4.29 B). A difference of -4 means that the gene showed it's maximal expression 4 hours earlier under constant light condition. Most of the genes had a negative difference. In addition the phase shifts of the genes within one CT group was calculated and the results are plotted as a surface diagram (Figure 4.29 A). This diagram confirmed the assumption that genes reached their maximal expression in the course of a day earlier under constant light condition.

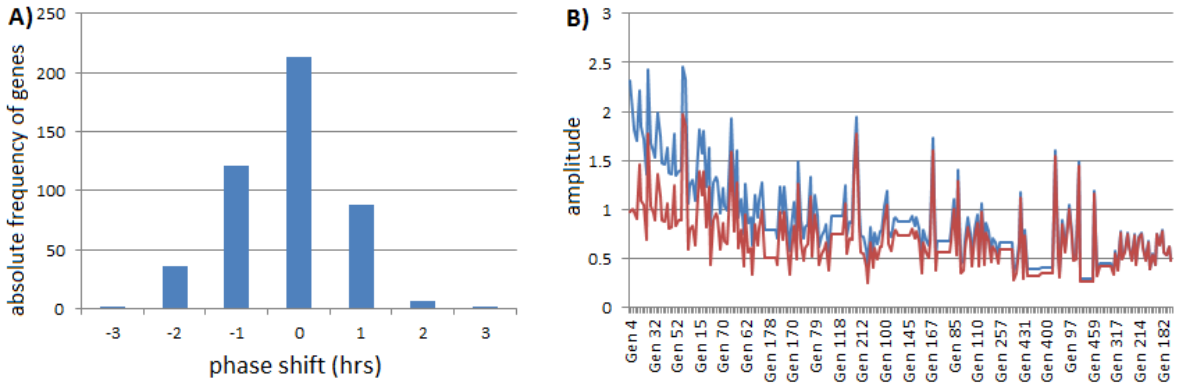

Figure 4.30 Changing parameter values of the first and second day under LL condition

A) Comparison of the period of rhythmically expressed genes. The shift was calculated using the period estimated for the genes of the first day in LL condition as a reference. B) Changing amplitude values for the first (blue) and second (red) day in LL conditions

For further comparison, the 48 h time course series under LL condition was disjoint into the first and second day. It has to be mentioned that the time-series of the first day is not

equal to the first 24 h in constant light condition. The measuring started 28 h after transferring the plants from LD to LL condition. To receive time- series with 12 time-points in a 4 h interval we simply duplicated the data and concatenated the time-series to a two day time course. At first we examined the change of period and amplitude values between these both days. The same trend was detectable: period and amplitude values decrease over time (Figure 4.30).

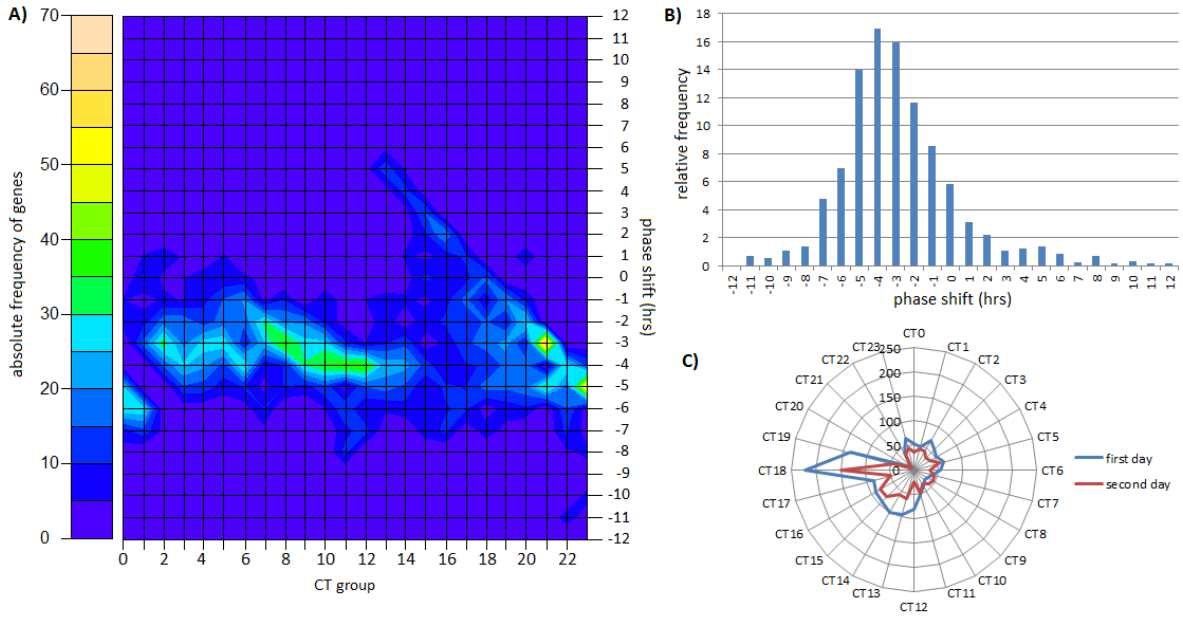

Figure 4.31 Comparison of the first and second day under constant light

A) Time shift topology graph plots absolute number of genes CT group shifted per time bin (y-axis) by the reference CT group (x-axis). A positive CT group shift reflects a later CT group than the reference condition and a negative one reflects an earlier CT group than the reference condition. First day in LL condition was taken as reference. B) Relative values of rhythmically expressed genes with the same time shift compared the first and second day of the time-series achieved under constant light condition. CT group classification of the first day was taken as a reference set. Difference was calculated by substituting the values of the second day from the first day within an interval from -12 to 12 in which CT0 was set as 0 C) Absolute number of genes per CT interval for the first and second day

However, damping of oscillation process was not as obvious as in LD to LL comparison, but still the effect was notable. Further we investigated whether and how the peaking time of

the genes changes and whether a pattern emerged. The distribution of the CT groups for the first and second day looks the same (Figure 4.31 C). Once again the difference between the CT groups was calculated. The CT groups determined for the genes of the first day served as reference set. Most of the genes had a negative shift which means the genes peaked on the second day at an earlier stage (Figure 4.31 B). Only in some rare cases the genes behaved in the opposite way and the expression of a small group of genes didn't change at all. The time shift for every single CT group was also estimated and confirmed the results from Figure 4.29 A (Figure 4.31 A).

5.1 Discussion

Normalization

We used kinetic microarray technology to examine the expression of different genes in one experiment simultaneously. The benefit of such a technology is the possibility to discover relationships and dependencies much easier. Commonly, the importance of data pre-processing in the data mining process is disobeyed, despite the fact that especially the normalization is necessary to correct expression values and to make various arrays comparable. The choice of the normalization technique influences the significance of the results. The more successful the method, the more reliable are the results. Until now, we do not have a “gold standard” and thus we browsed through the literature to search for the most adequate method. Finally, we decided to use 0.75- quantile normalization technique, because it’s a commonly used method in the field of microarray analysis and made it easier to compare our results to other published data. Another possibility was to use so-called “housekeeping genes” but we rejected this method. One reason was the general inaccuracy of the method, because it cannot be said exactly, whether there exist genes which expression is independent of environmental changes. The other reason was that we tested manually the known housekeeping genes for *N.attenuata* and we could not exclude the possibility that the gene expression differs according to the condition.

The visualization of the expression values of our replicates in boxplots (Figure 4.1) provides a first indication of distribution and variance in the data. To facilitate the interpretation of the visualized the data was logarithmized. It was noticeable that the number and distribution of outliers varies significantly among the replicates. Additionally it was found that the data shows a positive skew. A normal distribution has no skew, is perfectly symmetrical and the mean is exactly at the peak. The boxplot indicates that our data is skewed to the right as well as the mean is on the right of the peak value. Normalization stretches one tail of the distribution and shrinks the other, therefore the skewness is

removed (Figure 4.2) [9]. The performed normalization resulted in the desired effect and made the replicates comparable.

Experimental data

To unveil the genome of *N. attenuata* two time-series with three biological replicates were available. Due to the high material consumption, for each time-point another plant of the population was sampled. Thus, the time-series were pooled and assembled randomly by measurements of various individuals. Due to this "pooling" procedure the time-points may be regarded as independent. Hence, the time-series are per definition, even if the plants come from an inbred line, no true replicates. A first graphical visualization of the experimental data provides an overview of the data distribution. One the one hand, the boxplots displayed no large differences among the replicates on the other hand, they are only representative for single time-points of the time-series. The number of outliers as well as the distribution differs slightly, which is only normal, considering the large biological variation even among individuals. This is precisely the reason why biological replicates are so important. Even organisms with the same genetic constitution are not necessarily 100% identical due to stochastic fluctuations in gene expression. It is assumed that every nucleus is not deterministic in its expression repertoire and therefore the process of gene expression is not determined completely [45]. Furthermore it can be assumed, that no two cells act in the same way due to the large number of cells per tissue. Based on these assumptions we first attached no great importance to the observed differences among the replicates. The performed normalization resulted in the desired effect and made the replicates comparable. Normalization procedure was followed by the analysis with ARSER, which reveals glaring differences among the biological replicates relating to the number of ascertained oscillating genes. This becomes particularly obvious if we focus on the second replicate under LL condition. The resulting number of rhythmically expressed genes was triply as high as for the other two replicates and exceeds even the determined number for

replicate 2 and 3 under diurnal light conditions. For *A. thaliana* the literature reveals a value of 2-4% of genes oscillating with a circadian rhythm and 6-15% of genes showing a diurnal expression pattern [56, 69]. These data suggest that the results for replicate 2 in LL condition contain a high number of false positives. The boxplots didn't show any serious differences among the replicates and a confusion could be excluded within the realms of possibilities. By checking the internal standards on the chip a technical error could also be excluded as far as possible. In consideration of the fact that no direct cause or error could be found for replicate 2 (LL condition) and an exclusion of the replicate would reduce statistical significance we decided that this replicate will also be part of further analysis. It is an accepted procedure to average the expression values during data preprocessing, but this was inappropriate for our purpose. The mean is particularly susceptible to the influence of outliers and we could not identify the reason why the number of oscillating genes is so different for the replicates under LL conditions. Therefore, a time-series generated of mean expression values was not considered, because this time-series would not be representative and could lead to strange results. Under LD conditions replicate 1 stood out above all other replicates and once again we decided not to use an average expression profile. There was a risk that the mean value generates a time-series that would represent the real gene expression in an inadequate manner. Due to the influence of outliers the average time-series could give a false impression of the gene expression. The consequence would be that imprecise conclusions are drawn concerning circadian-regulated genes. Commonly, it is assumed to infer the probable gene expression profile of an average cell in a population from an average time-series. In truth, an averaged cell is obtained which per se does not exist [45].

Although the absolute number of oscillating genes differs widely among the replicates, the Venn diagram (Figure 4.6) shows a strong overlap, indicating that the same genes were detected. We assumed that the genes in the overlap showed a rhythmic expression profile and determined the ZT and CT groups. The subsequent visualization revealed that the single replicates show a completely different distribution pattern (Figure 4.2). The ZT

group distribution did not only differ by a few hours, but the peaks can be found in completely different parts of the day. Thus, entirely different functions could be assigned to rhythmically expressed genes of each replicate. Most of the oscillating genes of replicate 1 reached their maximal expression shortly after sunrise, indicating a probable participation of the genes in the light respond reaction. In the middle of the day (ZT6), when the sun is at its maximum, most of the rhythmically expressed genes of replicate 2 reach their maximum expression. For instance indicating a possible participation in photosynthetic process. However, a large part of oscillating genes of replicate 3 peaked in the night (ZT 18 and 19). These genes may participate in the diurnal metabolism of starch, because it is known that starch is degraded during the night [49, 88]. Detailed knowledge about the function of genes could not be obtained because an imprecise gene annotation prevented a gene enrichment analysis in the set period of time. It can thus only be speculated about the involvement of the genes in metabolic pathways. Here it should be pointed out that the replicates, considered individually, allow completely different conclusions about the involvement of the circadian clock in metabolic processes. In the CT group distribution of the replicates in constant light conditions differences were also clearly visible. Due to the significantly higher number of oscillating genes classified within replicate 2 a simultaneous representation of the distributions for all replicates does not allow a differentiated evaluation. Most of the genes, about 60%, of the second replicate were assigned to CT group 21. This enormous peak superimposed the distribution pattern of replicate 1 and 3, whereby a separate display was necessary for these two distributions (Figure 4.3). These two distributions show a clear overlay, the respective peak (the CT group, that is assigned to most of the genes) is exactly 12 h apart. Most of the genes of replicate 1 peaked right after subjective dawn (CT1), whereas the genes of the third replicate peaked right after subjective dusk (CT13). Metabolic processes, probably regulated by the circadian clock, thus seem to be completely divers. It is striking that both distributions show a smaller peak at CT 18, suggesting that the internal clock exerts a high impact on the genes of this CT group. It must be born in mind that an overlap in the

distributions does not automatically mean that the same genes were assigned to this group for each replicate.

In the analysis of biological replicates was found that these are not particularly good for accurate analysis of the genome of *N.attenuata* due to their large variance. Since the exact cause for the differences could not be determined, also an exclusion of individual replicates was waived. Instead, the generation of new time-series by random selection of the expression values of biological replicates was adopted.

Randomization

Shedden and Cooper [71] statistically reanalyzed microarray data of gene expression in human cells after double-thymidine block synchronization and were able to refute the original conclusions. The original microarray data was presented to support the existence of oscillating gene expression in human cells. To test whether the level of cyclic expression could also be explained by random fluctuations such as biological or technical noise they randomized the data. In contrast to our procedure they randomly selected the expression values for a given gene across all time-points. In other words, they conducted a random permutation test in such a way that any permutation had the same probability to occur. Even though the time-points of our time-series were independent of each other, due to the pooling, we strived to keep the temporal order. Our aim was not to simulate random fluctuations, like *Shedden and Cooper* intended, but to guarantee that on average no systemic differences in covariates emerge. Although our method of randomization should be distinguished from common randomization test, similar conclusions were reached. The analysis of our random time-series reveals that the absolute number of oscillating genes was nearly the same for all time-series and the randomization process leads to a more balanced pattern (Figure 4.8). *Shedden and Cooper* were able to demonstrate that the rhythmic gene expression observed in human cells could be explained as results of random fluctuations and therefore any attempt to identify oscillating genes would be contaminated with a large number of false-positives. Indeed,

we found that the lists of genes classified as rhythmically expressed in the various time-series differ widely in their composition. We calculated the overlap of an increasing number of compared datasets and the result showed that the number of oscillating genes is rapidly reduced in the intersection of 2~8 datasets (Figure 4.9). To prevent an organism or algorithm specific effect we repeated the analysis for another higher plant, *A. thaliana*, and with HAYSTACK, another algorithm to detect rhythmically expressed genes. The result was the same indicating that the results of single time-series contain a high number of false positives. Therefore, we decided to classify a gene as rhythmically expressed if it is identified in at least 20 out of 30 time-series. This threshold was chosen on the basis of Figure 4.9.

Yang and Su [86] compared the performance of their own algorithm (ARSER) with two widely-used rhythmicity detection techniques (COSOPT and Fischer's *G*-test) and could show that their method is considerably more accurate [86]. We analyzed our randomized time-series with ARSER as well as HAYSTACK and found that ARSER detects significantly more genes and nearly all genes identified by HAYSTACK. This result suggests two possible interpretations: first, HAYSTACK is more stringent and the results are more accurate and second, ARSER is more sensitive and able to detect even slight periodicities in gene expression profiles. Through the detailed analysis of genes detected only by ARSER we were able to confirm the second assumption (Figure 4.11).

Different intervals for ZT/CT group determination

The main object of this thesis was to investigate the gene expression under different conditions. Additionally various tissues should also be examined. To allow comparison among the different datasets, it was necessary to find a suitable classification of the genes in various groups according to their expression profiles. The literature research revealed the opportunity of two main principles: a scale-free clustering based on distance metrics or distribution models and an interval scale classification. The advantage of a clustering method is the efficient grouping of genes with highly similar expression

profiles. Hence, this classification seems to be very useful to detect functional correlations among genes and it is very likely that the returned clusters contain genes with similar functions. Such a grouping facilitates for example a reconstruction of a metabolic network. Beside this immense advantage the clustering method was inappropriate for our purpose. To accurately estimate the gene expression under different conditions and to detect and quantify changes an equidistant scale is necessary. We decided to use the ZT/CT group definition of *Ueda et al.* [82] and a 1 h interval to assign genes to different ZT/CT groups based on their molecular peaking time. Upon closer examination of the definition, we discovered a significant drawback: the borders of the groups are fixed. We assumed that this fact could lead to some distortions of the results. Genes, that peak nearly at the same time, will be assigned to different ZT/CT groups under various circumstances and functional relationships might be lost. However, we could show that a shift of boundaries had no effect on the ZT/CT group distribution (Figure 4.16). While samples of the microarray experiment were collected in a 4 h interval the chosen equidistant scale has a one hour interval, may causing inaccuracies in the assignment. Therefore, we tested different intervals ranging from 1 to 4 hours. As expected, the enlargement of range size resulted in a higher number of assigned genes per group (Figure 4.17). We expected only a slight change of gene expression under different conditions and therefore such large intervals would prevent a sophisticated analysis. On the basis of these results and considerations we decided to use the ZT/CT group definition of *Ueda et al.* without any changes.

Simulated data

As has already been pointed out, we decided to combine the biological replicates in a randomized manner to unveil the genome of *N.attenuata*. The number of oscillating genes differs widely among the biological replicates, whereas the results of the randomized time-series were balanced (Figure 4.8). Additionally, all time-series showed a similar ZT and CT group distribution pattern. We decided not to rely on the results of single

time-series, but to classify a gene as rhythmically expressed, if it was identified in at least 20 out of 30 time-series. This choice was made based on Figure 4.9, which indicates that once this number of compared datasets is exceeded, the absolute frequency of oscillating genes does not change significantly anymore. To corroborate a belief, simulated data was generated. The exact number of circadian-regulated genes is largely unknown for higher plants, like *A. thaliana* and *N. attenuata*. Therefore it was difficult to confirm our results and methodology. In simulations the exact number of periodic and non-periodic patterns is known. Hence, we applied different performance evaluation measures. In effect, the determination of oscillating genes is a binary classification. There are only two possible outcomes: either a gene is rhythmically expressed or not. The accuracy of this classification can be estimated by a confusion matrix [44]. We could demonstrate that a threshold of 20 led to exclusively correct classifications. Figure 4.12 confirms our assumption that an increasing number of compared datasets decreases the number of false positives whereas the number of true positives does not change significantly. Starting from a threshold of 20 not a single false positive was detected and the intersection contained only periodic patterns. In order to prevent that the results were algorithm specific we tested the same data with HAYSTACK. These results confirmed the effect observed within the results received by ARSER to its full extent. Here also, an increasing number of compared datasets led to a decrease in the number of false positives. Extrapolated to the experimental data we can conclude that the genes identified as rhythmically expressed in at least 20 out of 30 time-series are oscillating genes. An oscillating amount of RNA in the course of a day is a commonly accepted evidence to classify a gene as circadian-regulated [85]. Thus, we could presume that the genes of LD_20 and LL_20 are really influenced by an endogenous clock. Secondly, it is considered that the number of oscillating genes within a genome is overestimated due to inaccurate models [20]. The results of the simulated data suggest yet a contrary assumption. For each examined dataset the sensitivity, thus the true positive rate, never exceeded a value of 0.66. Therefore, only 2/3 periodic patterns were detected of the overall number. 33% of

rhythmic expression profiles remained undetected. Naturally an examination of the results received by single time-series displayed a higher number of true positives, however, also more false positives were reported. To take these results into account a threshold of 20 should not be laid down, but should be customized to the specific requirements and aims of new experiments. If the focus is on determining the number of oscillating genes as error-free as possible a threshold of 14, ideally 20, should be picked. In rare cases, the number of true positives exceeds the false positives in a matter of importance and here, a threshold between 2 and 6 should be selected. Within this range the number of true positives was highest and more genes are considered. Likewise remarkable is the fact that a sensitivity of only 0.33 was reported for the third dataset. This dataset contained only 5% of periodic patterns. The sensitivity did not change even when the threshold was altered. Therefore, the value seems to be independent of the threshold. The relatively small number of true positives was due to the sensitivity of the algorithms. ARSER as well as HAYSTACK had issues in recognizing the periodic patterns. Another explanation could be that the randomization technique limits the number of true positives. The precise reason could not be identified. In conclusion, the accuracy of the randomization and classification process seems to be independent of the ratio of periodic patterns within the time-series. Furthermore, a threshold of 20 always leads to a list of genes without any false positives. Nevertheless, due to the sensitivity of the algorithms the number of oscillating genes will be underestimated if the overall number of periodic patterns within the time-series is small. To solve this optimization problem new algorithms have to be invented to increase the overall sensitivity.

To average the expression values of biological replicates during data preprocessing is a commonly accepted procedure in scientific community [40]. Comparisons made between the mean and our method (threshold=20) yielded several interesting results, showing that our method performed better and provided the more reliable results (Figure 4.15). Whereas the sensitivity was nearly the same for both methods, the average method reported more false positives. While for our method the specificity stayed constant at 1,

indicating that no false positive results were included. For instance, if we focus on the first data set a specificity of 0.93 was reported for the average method. At first glance, the distinction of 0.07 seems to be less significant, but it proposes that the average method detected 1856 false positives, while our method detected non. The significance of the differences becomes apparent when the precision of both methods is compared. Our method always reached a value of 1 whereas the average method reached a maximum value of 0.63. The results indicate that the average method reported almost as much true positives as false positives. Our methodology fully delivered relevant records and led to the more reliable results. It can therefore be assumed that the procedure to average the expression values into a mean provides inaccurate results. Exploring the reasons it is pertinent to look at the publication of *Levsky and Singer* discussing the „myth of the average cell"[45]. The main statement of their publication is that there is no average cell. Gene expression occurs within single cells and although this assumption is commonly accepted in the scientific community it is often disobeyed. Instead, the information comes from samples containing millions of cells. On the one hand it is much easier to examine large samples than single cells, on the other hand, it is assumed that this mixture of cells represents biological variation. What we obtain is not the state of an average cell within a population, but an "*averaged"* cell [45]. Just like the average cell does not exist there is no average time-series either. The importance of biological replicates is not up for discussion [8, 62, 87], but the use of the average expression profiles ignores the natural variance among the individuals. Conclusions drawn from average time-series leads to inadequate general statements. Assumptions and networks, based on these conclusions only give a distorted picture of reality. Each cell reacts in a non-deterministic manner to the same external stimuli and starts an individual transcription program. Correlated transcription in individual cells might occur by accident. The same assumption is true at the level of organisms. It is quite natural that biological replicates differ in their expression profiles and the captured variation should not be destroyed by using the mean. The average method leads to time-series which does not exist in nature. Each time-point of a time-

series is represented by an expression value which does not reflect reality. In the overall impression, the randomization technique generates also time-series, as they were not recorded in this composition, but the individual time-points represent real values. The applied pooling method for each replicate provides the basis for the independence of each time-point. Our methodology reinforces the randomization and repeats only this effect. The generated time-series were much more in line with real expression profiles, as the data of the average method.

Cluster analysis

The dataset from *Kim et al.* [42] was used to investigate the gene expression within different tissues. These time series contained only six time-points. ARSER can be applied only to time course series with at least eight time-points (Figure 4.16). Therefore, it was necessary to develop another method to determine the number of oscillating genes and the ZT groups. We decided to use a method similar to the proposed algorithm for clustering short time-series expression data from *Ernst et al.* [18]. For each ZT group we generated a periodic model profile based on the parameters values returned by ARSER for the oscillating genes in 12 h light and 12 h dark cycles. Further we calculated the correlation coefficient between the model profiles and the experimental data. Each gene was assigned to the profile with the highest correlation coefficient. When we tested our profiles we realized that an accurate differentiation among neighboring ZT groups is nearly impossible due to only small phase shifts in molecular peaking time. The correlation coefficients of these groups were very similar. It should, however, been seen in the context that a sampling interval of 4 h was used. Such an interval makes it difficult to receive a 1 h resolution. Furthermore, the revision of the profiles leads to the conclusion that a threshold for the correlation coefficient of at least 0.7 is necessary, to distinguish among the various ZT groups. A value below 0.05 would lead to a broad spectrum of eligible ZT groups. Likewise, it was clearly evident that a threshold above 0.9 would only lead to inadequate results. Expression profiles of genes which belong to the same ZT

group are very similar, but still not identical. The reasons for differences found among members of the same group are due to biological variations and noise that bias gene expression measurements. Consequently, even within one ZT group a correlation coefficient of 1 is impossible, but should be above 0.95. Hence, if the correlation coefficient is set too high, only a very small amount of genes would be assigned to specific ZT groups. We tested our clustering method on the LD 12:12 dataset to define the best threshold. As reference we used the ZT group distribution received with the molecular peaking time method. As could have been expected, the number of genes assigned to model profiles increases with a decreasing threshold. The lower the threshold value, the more variance was tolerated. At the same time the number of false positives was increased. A threshold for the correlation coefficient of 0.9 leads to the most reliable results, simultaneously the absolute number of assigned genes was decreased. We were not interested in quantity but rather in quality. Therefore, a threshold of 0.9 was chosen. To take into account that the correlation coefficients of neighboring ZT groups are very similar the algorithm always returns three ZT groups with the highest coefficient. Using the ZT groups determined with molecular peaking time method as reference we could show that our clustering algorithm is really sensitive. Therefore, our clustering algorithm is able to determine ZT groups with an accuracy of +-1.

Comparison LD 12:12 and LD 16:8

We investigated the impact of a change of day-length on the phase. The results showed that the genes of plants under long-day conditions reached their maximal expression at a later time-point, indicating that its phase responds to the time of dusk. In adaptation to the diurnal rhythm, the endogenous oscillator shows separate bouts of morning and evening activity to synchronize the metabolism with the environment. These two feedback loops are intracellularly coupled by cell-cell signaling [75]. The analysis of the data of two photo-cycles revealed a difference concerning the responds of morning and evening gene expression. As expected, most of the evening genes reached their

maximal expression with a delay of 1 to 3 hours. With this result, we could confirm the statements from the literature for Arabidopsis [56]. Plants, which are sessile organisms, cannot avoid unfavorable environmental conditions, therefore it is important to adapt to changes as quickly as possible, in order to achieve optimal profit from the given resources. For instance, in this case the plants reacted to the elongated day-length by shifting the evening gene expression to the time of dusk, to use the full capacity of daylight. It has been proven that a perfect synchronization of metabolism with the environment leads to a selective advantage [16]. The plants grow faster due to an enhanced chlorophyll content and photosynthetic carbon fixation rate [16]. Therefore, a correctly tuned circadian system, increase the fitness of a plant and improve survival. Likewise, we expected an advanced gene expression of the morning genes. Surprisingly the molecular peaking time of the morning genes was not consistent. In long-day conditions the night is shortened by 4 h. Therefore, one would expect that the genes of the morning-expressed loop peaked at a former time-point, to anticipate the dawn. Some of the morning genes showed exactly that kind of respond and reached their maximal expression 1 to 2 h in advance, compared to the LD 12:12 conditions. However, the peaking time of some of these genes was delayed by 2 to 4 h, indicating that there was no respond to the time of dawn. This observation would also indicate that the feedback loop of the morning and evening oscillator has a higher impact on the synchronization process then the environmental cues. We hypothesized that time-delayed morning genes are synchronized by internal signals. To test this assumption we screened our LL_20 list to check whether the morning genes were specified here. If the test is positive, one might assume that the expression of these genes is controlled by the internal clock. Our result could either confirm or refute the hypothesis. The identification of the internal effect is not very specific and informative, for which reason results can only be interpreted with difficulty. Based on the area-proportional Venn diagram a small number of morning (and also evening) genes seems to be regulated by the circadian clock, because they kept their rhythm under LL conditions. This general statement is commonly known, but we cannot go into the matter

any further. To do so a gene enrichment analysis as well as a detailed investigation of the morning and evening genes under LL conditions is necessary. The question, whether the coupling of the oscillators or environmental cues are responsible for the observed phase shift remains open.

Comparison leaf LD 16:8 and root LD 16:8

We examined the gene expression in different tissues to test the assumption that the circadian clock is organ-specific. To ensure optimal fitness and adaption to the environment an appropriate synchronization of the rhythmic outcome of internal oscillators in different organs is crucial [52]. In contrast to animals, plants do not own a central nervous system that controls the synchrony of endogenous clocks located in different tissues [80]. *James et al.* [36] could show that the circadian clock in shoots and root are synchronized by a photosynthesis-related signal from the shoots. With our experiments we could confirm most of their findings, although the results are not really comparable due to the different experimental conditions. Nevertheless, our results revealed that the internal clocks in the shoot and root tissue show an organ-specific behavior (Figure 4.25). For both tissues we determined a different number of oscillating genes and the area-proportional Venn diagram (Figure 4.24) shows that there is a large number of genes oscillating only in one of the two tissues. The function of these genes could not be estimated, but it is likely that they are involved in organ-specific metabolic networks. For instance, the genes of the shoot tissue are likely to participate in photosynthesis-related processes, light respond reactions or in the mechanisms of leaf movement. However, it is conceivable that the genes from the root tissue are involved in pathogen defense or in the production of metabolites which are then transported into the leaf tissue. Furthermore, the overlap of oscillating genes between both tissues is relative small. It has also been noted by *James et al.* that in roots the circadian clock controls the expression of only a restricted set of genes, indicating organ specificity of the clock [36].

The observation that the expression of common oscillating genes is delayed by 1-2 h in the root (Figure 4.25) leads to the suggestion that the internal clocks of the two tissues are coupled and not autonomous as it is proposed by *James et al.* For N. attenuata we could not examine whether the oscillators are coupled or not, because for the root tissue we had only gene expression data in LD conditions. Though, the only evidence to proof that rhythms are controlled by different oscillators is to demonstrate desynchronization of the rhythms under constant external conditions [33]. Indeed, *James et al.* could observe that some clock genes lose synchrony in roots and shoots under LL conditions [52]. Due to the lack of data we could only show, that the expression of most of the genes in the root is delayed compared to the shoots in LD condition. Thus, the gene expression seems to be phase shifted in both tissues.

Comparison LD and LL

At first we noticed that the average number of oscillating genes in LL conditions is much lower than under LD conditions. This result was not surprising, given that the endogenous clock is synchronized to the environment mostly by external signals. In LL condition one of the most effective Zeitgeber, the light-dark cycle which entrains the internal oscillator, is missing. As a consequence of the absence of external stimuli the oscillating system demonstrates a free-running period. We could observe that 8% of the genes, which were identified in LD conditions as rhythmically expressed, exhibit a persistent rhythm in LL conditions with a period close to 22-25 hours. We expected a shortened period, because in LL conditions the organisms demonstrate a natural period close to 24 h. Nevertheless, higher plants are multioscillatoy systems and in any population, and even within an organism, some oscillators will always be inherently faster or slower. Additionally *Hastings and Sweeney* [29] proposed that individuals show significant differences in natural period. This is precisely a point we have to take into account, because our random time-series were generated based on the expression values of three biological replicates. Secondary, each time-point was sampled from a different

individual. Thus the expression values of one time-series symbolized a mixed output of various oscillators. This effect was even intensified by the applied randomization procedure. We could not only detect a shortened period but also a decreasing amplitude. With the passage of time under LL condition, the waveform broadens, and the rhythmicity of the gene expression gradually damps out. Thereby we could demonstrate that two very important parameters of the oscillating process were influenced by the change of photo-cycles. The oscillators within a plant are intracellularly coupled by cell-cell signaling. However, the comparison of gene expression in root and shoot tissue demonstrates that at least a phase shift is possible. The various endogenous oscillators of individuals within a population are only weakly coupled. Nevertheless, the behavior of the whole population depends on the width of natural frequency distribution. If the width of the distribution exceeds the strength of coupling the oscillators are unable to synchronize even they start in unison. Due to the weak coupling and the lack of external cues, which reset the clock in LD condition every day, the oscillators drift out of phase. Incoherence and a cacophony of oscillators are the results. Additionally, a short-term effect occurred right after transformation from LD to LL. The phase is not reset immediately but comes to its stable position only after several cycles. The time which is needed to come back to a stable position depends on the organism and its complexity. Sample collection started 28 h after the transfer therefore it is safe to conclude that the transient effects are still notable. Desynchronization may not necessarily occur in all plants, but the coupling between the organisms is too weak and under constant light conditions can therefore be no synchronization between the different individuals.

To summarize the results here are the main findings: We could proof that our randomization technique leads to a reliable set of genes which seems to be circadian-regulated. The clustering algorithm we developed is capable to analyze time-series with only 6 time points, to detect oscillating genes and to assign these genes to the right ZT/CT group with an accuracy of ±1. *N.attenuata* responds with a shift in gene expression so that the metabolism is synchronized with the environment. The internal clock in shoots and

roots show an organ-specific behavior, but we cannot make any statements whether the oscillators are coupled or not. Furthermore, we could show that the oscillation is damped in LL conditions and it is very likely that the individual clocks lose synchrony.

5.2 Outlook

Based on the results of this work two aims can be framed for the future: Improving the computational efficiency of the developed clustering algorithm, and to reconstruct a metabolic network of the oscillating genes that seems to be circadian-regulated.

With the improvement of the algorithm emphasis should be placed on the increase of accuracy of ZT/CT group determination by adapting different parameters. Although in this study already different thresholds for the correlation coefficient were tested, it is probable to improve the accuracy of the result by optimizing this value. However, the ZT/CT group determination is much stronger influenced by the choice of representative model profiles. So far, these have been selected on a random basis for each CT group. In the future, a method should be developed to generate adequate model profiles specific for the gene expression of each ZT/CT group. Additionally, consideration could be given to use another correlation coefficient.

A central goal of systems biology is to represent metabolic networks by mathematical models. In the near future a current gene annotation will be available for *N.attenuata*, which allows the estimation of the metabolic function of genes identified as rhythmically expressed. Systems biology is commonly viewed as a multistep process, whereas the whole methodology is built around the virtuous cycle of data mining (Figure 5.1). The results of this study enumerate a reliable set of circadian-regulated genes which could be used as components to build up the network. Thereby, it is possible to check off the first step of systems biology procedure. This first step includes most of the data mining process. The next step will be to reconstruct and define the interactions among these circadian-regulated genes. The internal clock influences many different metabolic

pathways, thus it is very likely that the reconstruction does not lead to one large network, but several small ones. To reconstruct the network structure the gene annotation as well as knowledge from the literature should be used. Subsequently, the network function must be simulated in silico. This simulation should provide information about possible weaknesses of the model, thus the model can be optimized later on. To complete one cycle of data analysis it is obliged to validate the computed network properties by comparison with actual phenotypic observations. Figure summarizes the whole data mining process in a flow chart, where the red boxes represent the outstanding steps to complete the cycle.

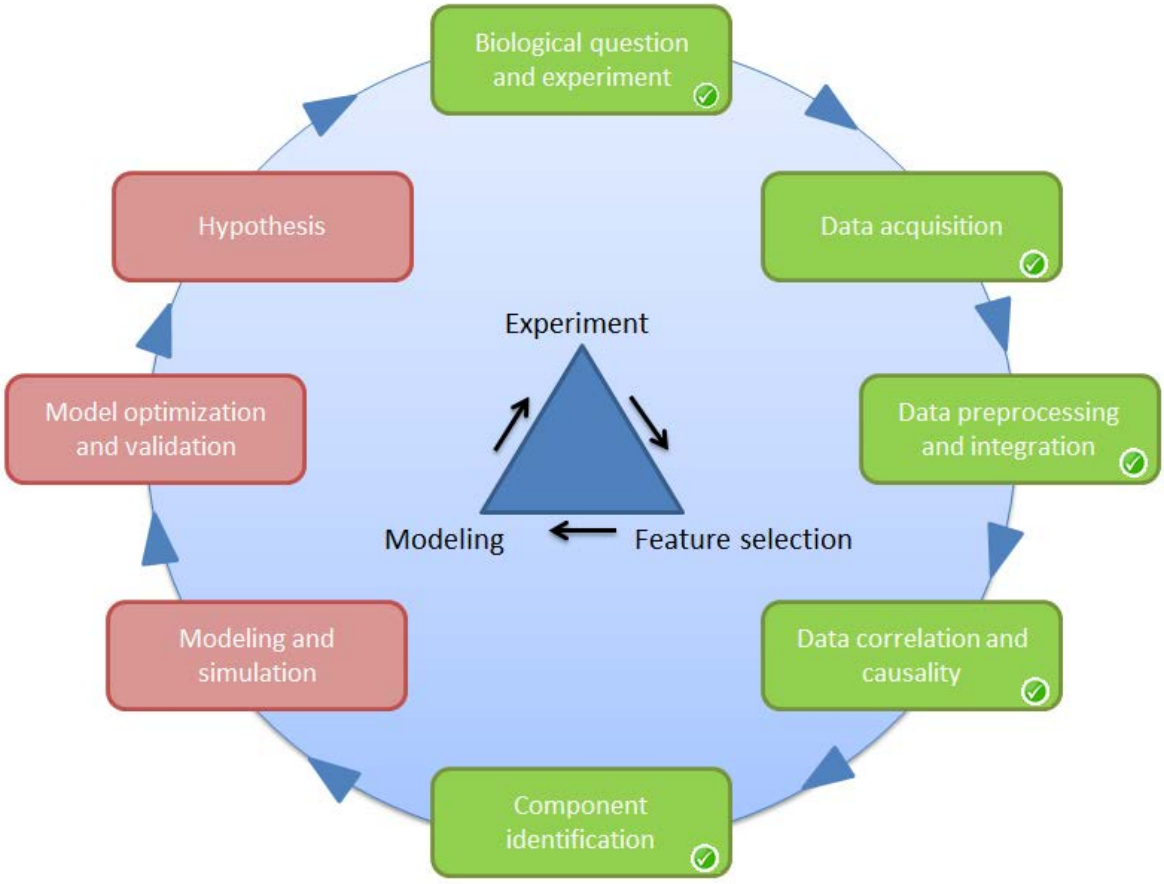

Figure 5.1 Flow chart of the whole data mining process
Displayed are the different steps of the data mining process, where the green boxes highlight the completed steps, and the red boxes represent the future work.

References

[1] Aschoff, J. (1966). "Circadian activity pattern with two peaks." Ecology: 657-662.

[2] Bar-Joseph, Z. (2004). "Analyzing time series gene expression data." Bioinformatics **20**: 2493-2503.

[3] Benjamini, Y. and Y. Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple hypothesis testing." Journal of the Royal Statistical Society **1**: 289-300.

[4] Bland, J. and D. Altman (1995). "Multiple significance tests: the Bonferroni method." BMJ **310**: 170.

[5] Bläsing, O., et al. (2005). "Sugars and Circadian Regulation Make Major Contributions to the Global Regulation of Diurnal Gene Expression in Arabidopsis[" The Plant Cell **17**: 3257-3281.

[6] Boes, T. and M. Neuhauser (2005). "Normalization for affymetrix genechips." Methods of information in medicine **44**(3): 414.

[7] Bünning, E. and K. Stern (1930). "Über die tagesperiodischen Bewegungen der Primärblätter von Phaseolus multiflorus. II. Die Bewegungen bei Thermo-konstanz." Berlin Deutsche Botanische Gesellschaft **48**: 227-252.

[8] Churchill, G. A. (2002). "Fundamentals of experimental design for cDNA microarrays." nature genetics **32**: 490-495.

[9] Cole, T. J. (1990). "The LMS method for constructing normalized growth standards." European journal of clinical nutrition **44**(1): 45-60.

[10] Consortium, M. "The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements." Nature Biotechnology **24**: 1151-1161.

[11] Consortium, T. G. O. (2001). "Creating the Gene Ontology Resource: Design and Implementation." Genome Research **11**: 1425-1433.

[12] Covington, M., et al. (2008). "Global transcriptome analysis reve als circadian regulation of key pathways in plant growth and development." Genome Biology **9**.

[13] Daan, S. (2010). A history of chronobiological concepts. The circadian clock, Springer**:** 1-35.

[14] Daan, S., et al. (2001). "Assembling a Clock for All Seasons: Are There M and E Oscillators in the Genes? ." Journal of Biological Rhythms **16**: 105-116.

[15] Danino, T., et al. (2010). "A synchronized quorum of genetic clocks." nature **463**: 326-330.

[16] Dodd, A., et al. (2005). "Plant Circadian Clocks Increase Photosynthesis, Growth, Survival, and Competitive Advantage." Science **309**: 630-633.

[17] Eldar, A. and M. Elowitz (2010). "Functional roles for noise in genetic circuits." nature **467**: 167-173.

[18] Ernst, J., et al. (2005). "Clustering short time series gene expression data." Bioinformatics **21**: i159-i168.

[19] Fukuda, H., et al. (2007). "Synchronization of plant circadian oscillators with a phase delay effect of the vein network." Physical review letters **99**.

[20] Futschik, M. and H. Herzel (2008). "Are we overestimating the number of cell-cycling genes? The impact of background models on time-series analysis." Bioinformatics **24**: 1063-1069.

[21] Gardner, M., et al. (2006). "How plants tell the time." Biochemical Journal **397**: 15-24.

[22] Genoud, T., et al. (2002). "Phytochrome signalling modulates the SA-perceptive pathway in Arabidopsis." The Plant Journal **31**(1): 87-95.

[23] Glass, L. (2001). "Synchronization and rhythmic processes in physiology." nature **410**(6825): 277-284.

[24] Goodspeed, D., et al. (2013). "Circadian control of jasmonates and salicylates: The clock role in plant defense." Plant Signaling & Behavior **8**.

[25] Haas, L. (1992). "Jean Nicot 1530-1600." Journal of Neurology, Neurosurgery & Psychiatry **55**: 430.

[26] Halberg, F. (1959). Withrow E (ed) Photoperiodism and related phenomena in plants and animals. Washington, AAAS.

[27] Halitschke, R. and I. Baldwin (2003). "Antisense LOX expression increases herbivore performance by decreasing defense responses and inhibiting growth-related transcriptional reorganization in Nicotiana attenuata." The Plant Journal **36**(6): 794-807.

[28] Harmer, S. (2009). "The Circadian System in Higher Plants." Annual Review of Plant Biology **60**: 357-377.

[29] Hastings, J. W. and B. M. Sweeney (1958). "A persistent diurnal rhythm of luminescence in Gonyaulax polyedra." The Biological Bulletin **115**(3): 440-458.

[30] Helfrich-Förster, C. (2009). "Does the morning and evening oscillator model fit better for flies or mice?" Journal of Biological Rhythms **24**(4): 259-270.

[31] Hennessey, T. L. and C. B. Field (1992). "Evidence of multiple circadian oscillators in bean plants." Journal of Biological Rhythms **7**(2): 105-113.

[32] Holmes, M. M., et al. (2004). "Adult hippocampal neurogenesis and voluntary running activity: Circadian and dose-dependent effects." Journal of neuroscience research **76**(2): 216-222.

[33] Honma, K. and T. Hiroshige (1978). "Internal synchronization among several circadian rhythms in rats under constant light." American Journal of Physiology-Regulatory, Integrative and Comparative Physiology **235**(5): R243-R249.

[34] Hughes, M., et al. (2010). "JTK_CYCLE: An Efficient Nonparametric Algorithm for Detecting Rhythmic Components in Genome-Scale Data Sets." Journal of Biological Rhythms **25**: 372-380.

[35] Hulsen, T., et al. (2008). "BioVenn–a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams." BMC genomics **9**(1): 488.

[36] James, A. B., et al. (2008). "The circadian clock in Arabidopsis roots is a simplified slave version of the clock in shoots." Science **322**(5909): 1832-1835.

[37] Jander, G. (2012). "Timely plant defenses protect against caterpillar herbivory." Proceedings of the National Academy of Sciences **109**(12): 4343-4344.

[38] Johnson, C. (2010). "Circadian clocks and cell division: What's the pacemaker?" Cell Cycle **9**(19): 3894-3903.

[39] Johnson, C., et al. (1998). A clockwork green: circadian programs in photosynthetic organisms. Biological Rhythms and Photoperiodism in Plants. P. Lumsden and A. Millar. Oxford, BIOS Scientific Publishers**:** 1-34.

[40] Kendziorski, C., et al. (2005). "On the utility of pooling biological samples in microarray experiments." Proceedings of the National Academy of Sciences of the United States of America **102**(12): 4252-4257.

[41] Kerwin, R. E., et al. (2011). "Network quantitative trait loci mapping of circadian clock outputs identifies metabolic pathway-to-clock linkages in Arabidopsis." The Plant Cell Online **23**(2): 471-485.

[42] Kim, S.-G., et al. (2010). "Tissue Specific Diurnal Rhythms of Metabolites and Their Regulation during Herbivore Atack in a Native Tobacco." PLoS ONE **6**.

[43] Kleinhoonte, A. (1929). "Über die durch das Licht regulierten autonomen Bewegungen der Canavalia-blätter." Archives Neerlandaises des Sciences Exactes et Naturelles **5**: 1-110.

[44] Kohavi, R. and F. Provost (1998). "Glossary of terms." Machine Learning **30**(2-3): 271-274.

[45] Levsky, J. M. and R. H. Singer (2003). "Gene expression and the myth of the average cell." Trends in cell biology **13**(1): 4-6.

[46] Lichtenberg, G. (1968). Schriften und Briefe. Fragmente, Entwürfe und Miszellaneen. W. Promics.

[47] López-Ochoa, L., et al. (2007). "Structural relationships between diverse cis-acting elements are critical for the functional properties of a rbcS minimal light regulatory unit " Journal of Experimental Botany **58**(15/16): 4379-4406.

[48] Lu, X., et al. (2004). "Statistical resyncronization and Bayesian detection of periodically expressed genes." Nucleic Acids Research **32**: 447-455.

[49] Lu, Y., et al. (2005). "Daylength and circadian effects on starch degradation and maltose metabolism." Plant Physiology **138**(4): 2280-2291.

[50] Luang, Y. and H. Li (2003). "Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data." Bioinformatics **20**: 332-339.

[51] Maere, S., et al. (2005). "BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks." Bioinformatics **21**: 3448-3449.

[52] Más, P. and M. J. Yanovsky (2009). "Time for circadian rhythms: plants get synchronized." Current opinion in plant biology **12**(5): 574-579.

[53] Matsuo, T., et al. (2006). "Real-time monitoring of chloroplast gene expression by a luciferase reporter: evidence for nuclear regulation of chloroplast circadian period." Molecular and cellular biology **26**(3): 863-870.

[54] McClung, C. R. (2006). "Plant circadian rhythms." The Plant Cell Online **18**(4): 792-803.

[55] Michael, T. P., et al. (2008). "A morning-specific phytohormone gene expression program underlying rhythmic plant growth." PLoS biology **6**(9): e225.

[56] Michael, T. P., et al. (2008). "Network discovery pipeline elucidates conserved time-of-day–specific cis-regulatory modules." PLoS genetics **4**(2): e14.

[57] Moore-Ede, M. C., et al. (1977). "Transient circadian internal desynchronization after light-dark phase shift in monkeys." American Journal of Physiology-Regulatory, Integrative and Comparative Physiology **232**(1): R31-R37.

[58] Morse, D., et al. (1994). "Different phase responses of the two circadian oscillators in Gonyaulax." Journal of Biological Rhythms **9**(3-4): 263-274.

[59] Naylor, S., et al. (2008). "Towards a systems level analysis of health and nutrition." Current opinion in biotechnology **19**(2): 100-109.

[60] Nomura, H., et al. (2012). "Chloroplast-mediated activation of plant immune signalling in Arabidopsis." Nature Communications **3**.

[61] Pittendrigh, C. S. and V. G. Bruce (1957). "An oscillator model for biological clocks." Rhythmic and synthetic processes in growth: 75-109.

[62] Quackenbush, J. (2002). "Microarray data normalization and transformation." nature genetics **32**: 496-501.

[63] Raj, A., et al. (2010). "Variability in gene expression underlies incomplete penetrance." nature **463**: 913-919.

[64] Roden, L. and R. Ingle (2009). "Lights, Rhythms, Infection: The Role of Light and the Circadian Clock in Determining the Outcome of Plant–Pathogen Interactions." The Plant Cell **21**: 2546-2552.

[65] Roenneberg, T., et al. (2008). "Modelling Biological Rhythms." Current Biology **18**: R826-R835.

[66] Roenneberg, T. and M. Merrow (2005). "Circadian clocks—the fall and rise of physiology." Nature Reviews Molecular Cell Biology **6**(12): 965-971.

[67] Roenneberg, T. and D. Morse (1993). "Two circadian oscillators in one cell." nature **362**(6418): 362-364.

[68] Rugnonea, M. L., et al. (2013). "LNK genes integrate light and clock signaling networks at the core of the Arabidopsis oscillator." PNAS.

[69] Schaffer, R., et al. (2001). "Microarray analysis of diurnal and circadian-regulated genes in Arabidopsis." The Plant Cell Online **13**(1): 113-123.

[70] Shannon, P., et al. (2003). "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks." Genome Research **13**: 2498-2504.

[71] Shedden, K. and S. Cooper (2002). "Analysis of cell-cycle-specific gene expression in human cells as determined by microarrays and double-thymidine block synchronization." Proceedings of the National Academy of Sciences **99**(7): 4379-4384.

[72] Slonim, D. and I. Yanai (2009). "Getting Started in Gene Expression Microarray Analysis." PloS Computational Biology **5**: e1000543.

[73] Somers, D. E. (1999). "The physiology and molecular bases of the plant circadian clock." Plant Physiology **121**(1): 9-20.

[74] Somers, D. E., et al. (1998). "The short-period mutant, toc1-1, alters circadian clock regulation of multiple outputs throughout development in Arabidopsis thaliana." Development **125**(3): 485-494.

[75] Stoleru, D., et al. (2004). "Coupled oscillators control morning and evening locomotor behaviour of Drosophila." nature **431**(7010): 862-868.

[76] Storey, J. and R. Tibshirani (2003). "Statistical significance for genomewide studies." PNAS **100**: 9440-9445.

[77] Streit, A. and R. Sommer (2010). "Random expression goes binary." nature **463**: 891-892.

[78] Takahashi, J. S., et al. (2001). Circadian clocks, Springer.

[79] Thain, S., et al. (2002). "The Circadian Clock That Controls Gene Expression in Arabidopsis Is Tissue Specific." Plant Physiology **130**: 102-110.

[80] Thain, S. C., et al. (2000). "Functional independence of circadian clocks that regulate plant gene expression." Current Biology **10**(16): 951-956.

[81] Turek, F. W. (1998). "Circadian rhythms." Hormone Research in Paediatrics **49**(3-4): 109-113.

[82] Ueda, H., et al. (2004). "Molecular-timetable methods for detection of body time and rhythm disorders from single-time-point genome-wide expression profiles." PNAS **101**: 11227-11232.

[83] Victor, A., et al. (2005). "cDNA-Microarrays – Strategien zur Bewältigung der Datenflut." Deutsches Ärzteblatt: A355-A360.

[84] Wichert, S., et al. (2003). "Identifying periodically expressed transcripts in microarray time series data." Bioinformatics **20**: 5-20.

[85] Wijnen, H., et al. (2006). "Control of daily transcript oscillations in Drosophila by light and the circadian clock." PLoS genetics **2**(3): e39.

[86] Yang, R. and Z. Su (2010). "Analyzing circadian expression data by harmonic regression based on autoregressive spectral estimation." Bioinformatics **26**: i168-i174.

[87] Yang, Y. H. and T. Speed (2002). "Design issues for cDNA microarray experiments." Nature Reviews Genetics **3**(8): 579-588.

[88] Zeeman, S., et al. (2007). "The diurnal metabolism of leaf starch." Biochem. J **401**: 13-28.

[89] Zeier, J., et al. (2004). "Light conditions influence specific defence responses in incompatible plant– pathogen interactions: uncoupling systemic resistance from salicylic acid and PR-1 accumulation." Planta **219**: 673-683.

[90] Zhao, L. P., et al. (2000). "Statistical modeling of large microarray data sets to identify stimulus-response profiles." PNAS **98**: 5631-5636.

Appendix

Series GSE30287

Status	Public on Dec 29, 2011
Title	Tissue specific diurnal rhythm of transcripts and their regulation during herbivore attack in *Nicotiana attenuate*
Organism	Nicotiana attenuate
Experiment type	Expression profiling by array
Summary	Idenfication and characterization of oscillating transcripts after elicitation with oral secretions from the sepcialist herbivore, Manduca sexta larvae
Overall design	Source leaves, sink leaves and roots were collected every 4 h for one days.
Contributor(s)	Kim S, Gaquerel E, Gulati J, Baldwin IT
Citation(s)	Kim SG, Yon F, Gaquerel E, Gulati J et al. Tissue specific diurnal rhythms of metabolites and their regulation during herbivore attack in a native tobacco, Nicotiana attenuata. *PLoS One* 2011;6(10):e26214.

Series GSE3461

Status	Public on Dec 01, 2005
Title	Diurnal gene expression in Arabidopsis thaliana Col-0 rosette leaves
Organism	Arabidopsis thaliana
Experiment type	Expression profiling by array
Summary	How do the transcript levels of leaf-expressed genes change in a normal day-night cycle? The interest is in genes that are regulated by the circadian clock and the diurnal component (i.e. light, metabolite changes). Plants were grown on soil in a 12/12 h light/dark rythm at 20°C day and night. 5 weeks after germination the rosettes of the non-

	flowering plants were harvested, 15 plants per sample. Plants were harvested at 6 timepoints every 4 hours beginning with the end of the night (still in darkness).
Overall design	3 biological replicates of the diurnal time series (6 times) were analyzed that were separately grown
Contributor(s)	Bläsing OE, Stitt M
Citation(s)	Bläsing OE, Gibon Y, Günther M, Höhne M et al. Sugars an circadian regulation make major contributions to the global regulation of diurnal gene expression in Arabidopsis. Plant Cell 2005 Dec; 17(12):3257-81.