ROYAL HOLLOWAY

UNIVERSITY OF LONDON

Doctoral Thesis

Uncovering the dynamic architecture of circadian gene expression in plants

Author:

Sandra Paulina Smieszek

Supervisor:

Dr. Paul Devlin

A thesis submitted in fulfilment of the requirements
for the degree of Philosophiae Doctor (Ph.D.) in Biology

in the

Centre for Systems and Synthetic Biology, Royal Holloway, University of London

Biological Sciences Bourne Laboratories

February 2015

1. Reviewer: Professor David Westhead
2. Reviewer: Dr. Enrique Lopez

Day of the defense:

Declaration of Authorship

I, Sandra Paulina Smieszek, declare that this thesis titled, 'Uncovering the dynamic architecture of circadian gene expression in plants' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: Sandra P. Smieszek

Date: 02/03/2015

Abstract

Plants have adapted to most of the planet's ecosystems, offering an amazing range of phenotypes and behaviours. Unlocking the underlying mechanisms like the circadian clock provides novel insights into the molecular hardwiring of plants and how they interact with their environment, furnishing new opportunities to develop better and more sustainable crops. The circadian rhythm is a roughly 24-hour cycle in the biochemical, physiological, or behavioural processes of living entities, including both plants and animals. It plays a crucial role in homeostasis and adaptation; hence, in agriculture. The difficulty in elucidating the circadian clock is rooted in the notion that two independent factors contribute to controllability: the system's architecture (whose components interact with each other) and the dynamic rules that capture the time dependent interactions between components. We set out to test mathematically the hypothesis that global circadian patterns of plant gene expression may be explained by progressive combinations of multiple promoter elements acting together. We proposed that the net effect of transcription factors acting at these elements could be responsible for the full range of phases observed in circadian output genes. We developed novel methods for the identification of circadian genes from short time-course microarray data and for the identification of the individual regulatory motifs which aggregate into coherent motif clusters capable of predicting the phase of a clock gene with high fidelity. We integrated gene expression profiles and protein interaction maps to provide a systematic and global view of combinatorial network modules underlying representative circadian programs. Furthermore, we integrate the newly discovered *cis* regulatory modules into the circadian regulatory network. Lastly, we developed circadian differentially inferred networks delineating the contrasting interactions between elements among genotypes. This study presents the analytical framework that should allow one to analyse the controllability of a complex system like the circadian clock in plants through the combination of driver nodes with their time dependent control reflecting the systems' dynamic logic. Such a circadian network provides a quantitative and holistic outlook upon a complex modular network of great agronomic importance.

Acknowledgements

This Thesis would not have been possible without the support of many of individuals whose suggestions have been instrumental during the years of my Ph.D. studies.

First and foremost I want to express my deepest gratitude to my advisor Paul Devlin. It has been an honor to be his Ph.D. student. He has showed me both consciously and unconsciously, the episteme and techne of being a scientist. I appreciate all his contributions of time, thoughts, and funding to make my Ph.D. experience productive and stimulating yet foremost his virtues, his intellect and patience as solely given the knowledge and time can one reach the desired aims. The passion he has for his research together with inquisitive mind was contagious and motivational for me, even during tough times in the Ph.D. pursuit. I am thankful for the excellent role model he has been and hence became my personal hero.

I am indebted to Dr Alberto Paccanaro and his entire network literati group, in the Centre for Systems and Synthetic Biology at the Department of Computer Science, who believed in me from the very beginning of my Ph.D. studies, and gave me the chance to pursue research and offered inevitable training and hours of dicussions. Without this it would not have been possible. He taught me the passion and the tools of research. I would like to thank a truly special scientist Dr Mikhail Soloviev, a person of great erudite, wit and passion. He has been a great mentor and his training will continue to radiate.

At same time I would like to thank my fellow group members who have been excellent role models, friends and a source of endless inspiration. They have greatly contributed to my personal and professional time. I am especially thankful to Dr. Haixuan Yang, Dr. Alfonso E. Romero, Dr. Prajwal Bhat, Dr Emilio Ferrara and Andrea Briones. I am particularly indebted with Dr. Yang, his suggestions helped me to improve the quality of this work. I am glad to have had the chance to work with Dr. Romero and Dr. Bhat in a number of research projects, and I am grateful to both of them for their precious help. I would like to acknowledge special group members Tamas Nepusz who has been a fellow early in my Ph.D. years, he became and will continue to be my great hero. I would like to thank all those who have come through the lab whose name is not mentioned nevertheless, You are my hero.

I gratefully acknowledge the funding sources that made my Ph.D. work possible. I was funded by the Crossland Scholarship.

I would like to thank mentors whom I had the pleasure to meet and who are the founders both in the field of molecular biology and bioinformatics, Professor Steve Kay

and Dr Altschul whom I had the pleasure to meet and those whom I have not met yet interacted with their highly extraordinary thoughts on paper, whose work has laid the foundation for my studies. As no man is an island and I am truly grateful to all those that have perpetuated my field of research and that keep on doing so.

Lastly, I would like to thank my family for all their love and encouragement. I am thankful for my family who raised me with a love of science and supported me in all my pursuits.

Yet most of all, I would like to thank my encouraging and patient husband Bart whos faithful support during all the stages of my Ph.D. was so appreciated.

To conclude and begin I wish to thank Mother Nature for providing us with the breathtaking beauty and conundrums we wish to solve

Thank you.

Contents

List of Figures

List of Tables

Dedicated to my Dearest Grandmother whose work I wish to continue and a true friend Apollo

Chapter 1

INTRODUCTION

”When we try to pick out anything by itself, we find it hitched to everything else in the Universe.”

— John Muir.

1.1 OVERVIEW OF THE PLANT CIRCADIAN CLOCK

The circadian clock is an endogenous 24 h timer, present throughout nature (McWatters and Devlin, 2011). This endogenous timekeeper can be found among all taxa of life and is reflected in biochemical, physiological, or behavioural processes. Plant circadian clocks generate stable oscillations with a period of approximately 24h that persist even under constant conditions, in the absence of any rhythmic environmental stimuli that impact the clock. Entrainment by environmental signals such as light and temperature can synchronize the clock to the period of the Earth’s rotation. Such synchronization confers a higher fitness to an organism as it allows one to anticipate daily cycles of light and temperature in a changing world. The clock fulfils a crucial role at the heart of cellular networks. Most organisms use circadian oscillators to coordinate physiological and developmental processes. The circadian clock plays a crucial role in plant biology and, hence, in agriculture because outputs controlled by the clock include the timing of germination, diurnal optimization of photosynthesis and floral transition (Nagel and Kay, 2012). Foremost circadian timekeeping in plants increases photosynthesis and productivity (Dodd et al., 2005, Ni et al., 2009, Yerushalmi et al., 2011). The machinery calls for production of certain proteins at the correct time, anticipating daylight early enough to produce the photosynthetic machinery, yet not too early so that certain unstable proteins start to degrade. In an experiment by Dodd et al ref, to demonstrate this leaf chlorophyll content, biomass and the carbon fixation were the assessed traits. Specifically when the clock was stopped in the model plant Arabidopsis thaliana, chlorophyll content was reduced under normal light/dark cycles Dodd et al. (2005). In the same study, when the aerial biomass was measured after several days, chlorophyll contenct was reduced by over 50 percent. In a different study it was shown that the *Arabidopsis* allotetraploids are larger and grow more effectively than the *Arabidopsis thaliana* and A. arenosa parents (Ni et al., 2009). The epigenetic modifications of the circadian clock genes of the hybrids and allopolyploids induced changes in the amplitude of gene expression in downstream genes involved in energy and metabolism, affecting chlorophyll synthesis and starch pathways. That suggests that the advantages gained from the novel genomic interactions led to altered control (that is resetting) of circadian clock-mediated processes which resulted in greater biomass.

The mechanisms by which the clock regulates productivity are numerous. The transcripts encoding the components of the Calvin cycle are circadian regulated and so are the transitory starch reserves managed by the circadian clock clearly showing the connections at different levels of the system to photosynthetic output (Graf et al., 2010). Plant cells communicate information about the time of day to their chloroplasts. The production of 'sigma factors' part of the cellular machinery in chloroplasts is under the control of the plant's clock, enabling the nuclear DNA to regulate activities of chloroplast genes (Noordally et al., 2013). Experiments indicate that individual nuclear encoded sigma factors communicate timing information to multiple chloroplast genes. Up to 70% of genes encoded by the chloroplast genome may be circadian regulated. It is probable that multiple sigma factors might impose circadian regulation of a single chloroplast gene, due to the functional redundancy of sigma factors yet different sigma factors affect different chloroplast genes depenedent upon the phase of the day(Noordally et al., 2013). Among other factors, circadian-gated light inputs, redox oscillations and the circadian $Ca2^{+}$ signals may also signal timing information to chloroplasts (Michael et al., 2008, Noordally et al., 2013). Growth itself is also clock regulated. Another output example is the growth of hypocotyls. Microarray studies have also been used to link several evening-phased clock proteins directly to output genes for example the *PIF4* (PHYTOCHROME-INTERACTING FACTOR4) and *PIF5* which encode bHLH transcription factors. Clock-regulation of *PIF4* and *PIF5* expression is essential for proper rhythmic growth, and their expression is under the control of the Evening Complex, which includes the LUX ARRHYTHMO, EARLY FLOWERING 3 and EARLY FLOWERING 4 respectively (*LUX*, *ELF3*, *ELF4*) a part of the central plant clock (Nozue et al., 2007, Nusinow, 2011).

Understanding how the circadian oscillator controls its output processes like photosynthesis and ,for example, increases productivity is a topic of great agronomic importance. Agricultural output can be increased thanks to both genetic modification and breeding that result in altered performance of the circadian clock as demonstrated for example through overexpression of B-box domain gene *BBX32* in *Glycine max* and hybrid vigour. Manipulating light signaling processes through affecting the expression/activities of these genes in plants can lead to improvedments in terms of agronomic characteristics, including, increased crop yield and improved stress tolerance. Altering the B-box genes may affects the levels of hexose sugars, altered levels of starch, and delayed senescence to name a few. It has been proposed that the beneficial phenotype is a result of changes in the timing of reproductive development. These could affect the duration of the pod and seed development period as proposed by Preuss et al.,. Constitutive expression leads to changes in the timing of the key phases of reproductive development via modulation of abundance of circadian clock genes during the dark light transition. It seems plausible that the model plant BBX32 modifies the input of light to the clock to result in a light dampening of clock rhythms near dawn resulting in lengthened reproductive phase. The discussed example is one of several output links with direct application to crop improvement. Specifically those relating to the control of photosynthesis will be discussed in the next chapters. Yet enhancing productivity through for example improvement in photosynthetic efficiency makes the understanding of the entire set of factors instrumental. These topics include: the negative feedback regulation of photosynthetic products and the regenerative capacity of the Calvin cycle, the optimization of canopy architecture and chlorophyll content, the reduction of photorespiration, the rate of recovery from photoprotective states, the engineering of Rubisco with increased carboxylation rates or with dramatically decreased oxygenase activity and the conversion of C3 plants to a C4 type system and the optimization of mineral nutrients availability and deployment to various parts of the plant and to cellular components in conjunction with circadian clock. The previously described examples were linked to productivity yet there are many more spanning other types of outputs. These processes span around the clock and include responses to abiotic and biotic stresses, metabolism and growth. The clock influences the responses to cold Espinoza et al. (2008), defence against pathogens (Goodspeed et al., 2013). To add another degree of relevance even treatment with glyphosate herbicide at different times of the day results in varying efficiencies, resulting from different light intensity and quality Sharkhuu et al. (2014). Glyphosate inhibits competitively the penultimate enzyme, 5-enolpyruvylshikimate 3-phosphate synthase, from the shikimate pathway Sharkhuu et al. (2014). A glyphosate resistant Arabidopsis knock out mutant has recently been identified displaying an elevated resistance phenotype in root/shoot growth. Sharkhuu et al., 2014 have recently shown that it is a mutation in phytochrome B that brings about altered expression of shikimate pathway genes and glyphosate resistance. Conversely, phytochrome B overexpression gives rise to a glyphosate-hypersensitive phenotype. Further links to the clock and PIF involvement were also suggestive of strong circadian regulation. More specifically it is likely that the mechanism involves negative regulation by the key clock gene, Circadian Clock Associated 1 (CCA1). There were pronounced differences in the expression of the shikimate pathway genes between the wild type and the phyB mutant. It became evident that it was the enhanced transcript accumulation which was the cause of the underlying resistance. Looking at the clock from the shikimate pathway angle alone gives another piece of evidence of how central the clock is to both the understanding of plant physiology and its applicability.

The plant clock has been studied among various plant species. The molecular biological study of the plant circadian clock began in 1985 with the cornerstone observation that the accumulation of three light-inducible transcripts, encoding a chlorophyll a/b binding protein, the small subunit of Rubisco, and an early light-induced protein oscillated in abundance in peas grown in 16 light dark cycles (Kloppstech, 1985). These oscillations persisted in plants transferred to continuous light. Another major breakthrough was presented in the seminal paper of Millar with firefly luciferase fusion analysis which became the mainstream mutant identification technique in the study of plant clocks (Millar et al., 1995). Studying the plant clock architectures raises speculation upon the origins of the clock. The ”flight from light” hypothesis for example posits an early evolutionary origin for clocks. This is consistent with the characterisation of a cyanobacterial clock. It has been shown that the oxidation-reduction cycles of peroxiredoxin proteins constitute a universal marker for circadian rhythms across all the domains of life (Edgar et al., 2012). Furthermore results suggest an intimate co-evolution of the circadian timekeeping after the Great Oxidation Event 2.5 billion years ago. To date angiosperm, bryophyte and green algal clocks have been studied all displaying certain degree of conservation. For example within angiosperms there is strong evidence for clock conservation in terms of architecture and function. That conservation has been shown in terms of individual genes and entire transcriptional clock regulated outputs. Whether there has been a one clock versus multiple clock origin within the eukaryotic lineage, it seems plausible to believe that to a great extent, logic in terms of transcriptional programmes has been preserved.

A series of circadian proteins act at specific times around the clock to reciprocally regulate the expression of a series of circadian genes, at the transcriptional and the post transcriptional levels. The model of the *Arabidopsis thaliana* circadian clock consists of a central loop and a series of peripheral loops. The central loop is termed a repressilator (a 3 component system), with three groups of clock proteins, each repressing expression of the next preceding component in turn to form a complete loop which oscillates with a 24 hour period (Huang et al., 2012). The 'repressilator' is a simple synthetic clock circuit consisting of a three-component negative feedback loop which provides a delayed negative feedback on all components and permits oscillations (Sprinzak and Elowitz, 2005). Essentially, the present understanding of the circadian clock in plants is mostly reliant upon transcription factors that mutually repress each other. In *Arabidopsis*, the first-identified clock genes were believed to function in a double negative feedback loop, with two morning-phased Myb-like transcription factors, CIRCADIAN CLOCK ASSOCIATED 1 (*CCA1*) and LATE ELONGATED HYPOCOTYL (*LHY*), repressing expression of an evening-phased pseudo-response regulator, TIMING OF CAB EXPRESSION 1 (*TOC1* aka *PRR1*), which in turn directly regulates the expression of CCA1 and LHY (Alabadí et al., 2001, Gendron et al., 2012, Huang et al., 2012). Specifically, the mechanism by which *TOC1* regulates *LHY* and *CCA1* transcription was elucidated not long ago with the *TOC1* protein shown to bind to the *LHY* and *CCA1* promoters via its conserved CCT domain. The acute effect of *TOC1* induction was to downregulate *LHY* and *CCA1* transcription. This showed that *TOC1* does not act to promote yet to repress expression of morning clock genes (Gendron et al., 2012, Huang et al., 2012). The mechanism for subsequent release of inhibition of *LHY* and *CCA1* transcription at dawn was revealed following the recent elucidation of the role of the evening complex within the clock. The expression of*LHY* and *CCA1* is followed by that of 4 PSEUDO-RESPONSE REGULATORS (PRRs), whose transcript levels peak sequentially starting with *PRR9* whose mRNA is expressed early at dawn. *PRR9* is followed by *PRR7*, *PRR5* and *PRR1* (*TOC1*). PRR protein levels also oscillate with a phase lagging slightly behind their respective transcripts (Fujiwara et al., 2008). The main function of the PRR proteins could be the maintenance of *LHY* and *CCA1* repression until the next dawn. However this raises the question of why it would be advantageous for the system to have several PRR genes expressed as multiple consecutive waves instead of a single PRR expressed as a broad wave. If these PRR proteins have both unique and overlapping target genes, as recently shown for *PRR5* and *TOC1*, sequential expression of the PRRs may represent a mechanism to initiate distinct output pathways within consecutive time-windows. Increasing the number of PRR genes may therefore raise the number of these windows, hence increasing temporal resolution within the circadian regulatory system (Carré and Veflingstad, 2013). The expression of *TOC1* towards the end of the day coincides with the expression of another transcription factor (*LUX*) and (*ELF3* and *ELF4*) (Carré and Veflingstad, 2013). The ELF3, ELF4 and LUX proteins assemble to form a protein complex called the evening complex (EC). *LUX* is the DNA-binding component of the complex and acts to recruit *ELF3* and *ELF4* to target promoters containing the target motif GATWCG with W representing A/T. The Evening Complex, in turn, indirectly activates *CCA1* and *LHY* by directly inhibiting the repressive PRRs. Quintessentially , the above connections are representing the core clock and do not cover the entire list of clock genes. The transcriptional network of the plant circadian clock is portrayed in Figure 1.1 which reflects the present state of knowledge.Furthermore *CCA1* and *LHY* act through the sequence known as the evening element (EE) (AAAATATCT) (Harmer, 2000). A *cis*-regulatory element named the evening element (EE) [(A)AAATATCT] has been found to be central to circadian clock function in plants. The elements sufficient for circadian expression include the dusk (Hormone Up at Dawn, HUD, CACATG), early-night (Evening Element, EE, AAAATATCT), and mid-night (Protein Box, PBX, ATGGGCC] and the Morning Element (ME, AACCACGAAAAT) (Covington et al., 2008, Harmer, 2000, Michael and McClung, 2002). Other motifs like the G-box, and GATA seem to play importance yet rather jointly in collaboration . They peak after dawn whereas proteins levels follow with a lag of about 2 hours.

The transcriptional regulation giving rise to the circadian rhythms in *Arabidopsis* has been furthermore depicted by Li et al., 2011, specifically highlighting the importance of both positive and negative stimuli giving rise to the circadian oscillations (Li et al., 2011a). It was shown that in *Arabidopsis thaliana*,FAR-RED ELONGATED HYPOCOTYL3 *FHY3*, FAR-RED IMPAIRED RESPONSE1 *FAR1* and ELONGATED HYPOCOTYL5 *HY5*, (which constitute the positive regulators of the phytochrome A signalling pathway), directly bind to the promoter of *ELF4* which is a component of the central oscillator, and activate its expression during the day, whereas the circadian-controlled *CCA1* and *LHY* proteins directly suppress *ELF4* expression periodically at dawn through physical interactions with these transcription-promoting factors (Li et al., 2011a). These results provide evidence that a set of light and circadian-regulated transcription factors act upon the *ELF4* promoter to regulate its cyclic expression, and establish a potential molecular link connecting the environmental light-dark cycle to the central oscillator. Furthermore it has been recently shown that the EC nighttime repressor plays a fundamental role in affecting the transcriptional modules by inspecting the warm-night and nighttime-light signals (Mizuno et al., 2014). This light-induced upregulation is dependent on phytochromes. Interestingly, the upregulation of these EC target genes is observed only light/temperature signals are concurrently fed into the repressor. This is another example of a seasonal timekeeping by the clock. The best characterized clock components repress expression of other core clock components. Recently a clock-regulated activator essential both for clock progression and control of clock outputs was depicted (Hsu et al., 2013). The latest work by Hsu et al. provides strong evidence that *RVE8*, a transcription factor that is similar to *CCA1* and *LHY*, regulates genes with peak expression in the early evening. Indeed, the circadian period of triple *rve4 rve6 rve8* mutants is four hours longer than that of wild-type samples. The evidence strongly suggests that *RVE8*, along with *RVE4* and *RVE6*, all play a key role in the circadian clock of *Arabidopsis* by switching on the afternoon/ early evening genes, which then induce in collaboration the expression of morning genes, thus starting the circadian cycle again (Hsu et al., 2013). Moreover the circadian clock seems to be responsible for the mechanism that generates the 12 hour rhythms through the interplay of circadian transcription factors (Westmark, 2014). The binding motifs for TF generating these rhythms are overrepresented in promoters of 12 hr genes. The generation of these rhythms conforms to the additive phase vector model proposed by (Ukai-tadenuma et al., 2011). The hierarchical network of circadian transcription factors which has been called the circadian transcriptional cascade and some branches of this model happen to generate the 12 hour cycles as proposed in the Westermark et al., study. It is now known that the circadian system exerts its effects at multiple levels for example through post-transcriptional regulation, posttranslational regulation, and chromatin remodeling introducing hierarchal levels of regulation. Some researchers stipulate that the non-transcriptional processes are the significant contributors to circadian rhythmicity (van Ooijen and Millar, 2012). A recent study has shown that PRMT5, a protein that transfers methyl groups onto subunits of the spliceosome is regulated by the light and dark cycle. This affects alternative splicing of some genes, making them potentially subject to circadian control (Sanchez et al., 2010).

Inputs to the clock are numerous. The main entraining stimuli that synchronize the internal clock with the external temporal environment are light and temperature. The dark to light transition at dawn is thought to be the main entrainment signal in plants, with phases change being dependant upon time at which it is applied (Devlin, 2002). The phase response curve shows phase advances prior to dawn, phase delays after dusk, invariance during the subjective day (Devlin, 2002). Light exerts a reciprocal influence upon the clock which in turn regulates aspects of light signalling. It is both the red-light sensing PHYTOCHROMES A, B, D and E, and the blue-light sensing CRYPTOCHROMES 1 and 2, which affect photoreception (Devlin and Kay, 2000). They cause a fluence rate dependant shortening of the period. It has been shown that the far red light absorbing form of phytochromes accelerates the clock pace whereas the red light absorbing form decreases the pace (Hu et al., 2014). Moreover, ELF4 has been noted to be involved in the light input. Recently it has been shown that the 3 regulators of the PHYA pipeline which include LONG HYPOCOTYL 5, Far Red Elongated Hypocotyl 3 and Far red Impaired Response 1 directly activate ELF4 during the day, elegantly linking the light signalling and the central oscillator cite Li. Despite the Q10 law which states that the rate of a biochemical reaction approximately doubles as temperature increases by 10C within a physiological range, circadian period is stable over a wide range of temperatures, a phenomenon highlighting the importance of the clock in homeostasis. It is suggested that alternative splicing might serve to adjust clock function in response to temperature changes. RNA-Seq analyses identified alternative splicing of numerous clock genes, and an event leading to the retention of an intron in CCA1 in a light and temperature-dependent manner was conserved across different plant species (Filichkin et al., 2010). RNA-seq offers by far greater depth of information for transcripts that are unannotated and indiscernible by microarrays such as newly annotated genes and splice variants. The first description of alternative splicing identified CCA1 transcripts which retain portions of intron 4 to produce two splice isoforms with premature termination codons (Filichkin et al., 2010). These variants were increased by high light and decreased in the cold. The functional consequence of CCA1 alternative splicing in the cold was shown more recently with the characterization of a third splice variant which upon translation retains all domains encoded by the full-length transcript (CCA1alpha) except the MYB DNA binding domain. This beta isoform causes interference through the formation of nonfunctional heterodimers and results in period-shortening of clock and output genes (similar to the cca1 lhy double loss-off function mutant) (Joon, 2012). Its production is suppressed in the cold. Essentially, it could be stipulated that the dynamic autoregulation by the splice variant of CCA1 coordinates the clock's cold acclimation. A systematic comparison of alternative splicing networks to the corresponding transcriptional programs should one day unravel the contribution of alternative splicing to the rhythms in transcript and protein abundance. Furthermore, it remains to be shown whether and how alternative histone modifications, such as phosphorylation, ubiquitination and sumoylation, also contribute to the plant circadian clock system. Additionally, a long list of hormones control the system in diverse ways. For example the cytokinins delay circadian phase, auxins regulate circadian amplitude and clock precision, and brassinosteroid and abscisic acid modulate circadian periodicity (Hanano et al., 2006) Finally, there are several other less well-undertood regulators of circadian oscillator function and these include cytosolic free Ca2+, cyclic adenosine diphosphate ribose, sugars, chromatin remodelling and protein phosphorylation cite Harmer Review.

The present model is insufficient and misses out on adequate activators to be rationally defined as complete. In essence, two component limit cycle oscillators can solely exist if one component is autocatalytic and negative feedback is present, with a limit cycle being the same repetitive cycle with no dampening (Bujdoso and Davis, 2013). As the clock in plant does not resemble that, one can believe it consists of 3 components and positive and negative arms, with repressors and activators being present (Hanano et al., 2006, Sprinzak and Elowitz, 2005). That opens novel avenues for research. Mathematical modelling has played an instrumental role in elucidating this central clock loop. To date at least seven models of the clock have been published. The present models, this is the kinetic model of Pokhilko et al., 2012, and the linear time invariant model of Herrero at al., 2012 clearly highlight the complexity of this system and did evolve on past learning.

They build on and modify an original Locke (2005a) et al., model. Mathematical models of the plant clock have shown that a combination of Hill, Michaelis-menten and linear functions can accurately reproduce gene expression dynamics observed experimentally. In such models, the degradation of gene products is assumed to follow michaelis-menten kinetics whereas the transcriptional activation and inhibition follows a Hill functional relationship, even though many proteins responsible are mostly unknown. The initial model of the oscillator has been revised and most notably all PRR proteins have been shown (biochemically and molecularly) to act in a repressive rather than a positive manner following the recent work of Gendron et al. (2012); and Huang et al. (2012). One could even stipulate the notion that, independently, mathematics is playing a leading role in unravelling the plant circadian oscillator. In historical order the models of the *Arabidopsis thaliana* plant clock included ”Locke 2005a”, ”Locke 2005b”, ”Locke/Zellinger 2006” , ”Kolmos 2009”, ”Pokhilko 2010”, ”Herrero 2012” and the present ”Pokhilko 2012” (Bujdoso and Davis, 2013). That progression shows a development that started with a simple positive/negative feedback model. Many important genes were then added to the circuitry starting out from that one core loop hypothesis. The mentioned models took into account the period shortening and lengthening behaviour of mutations in genes defined in these models. They were able to account for misexpression levels given certain mutant combinations. Shortly, experiments started confirming the presented predictions like the fact that that the triple mutant *cca1 lhy toc1* would be arrhythmic (Bujdoso and Davis, 2013). Throughout the discovery process, the current understanding of the clock did stand the test of time leaving space for modifications. There were indeed twists and turns such as the finding that the Early Flowering 4 (*ELF4*) gene (Doyle et al., 2002) was a core component had to be accounted for given that this mutation would cause the oscillator to stop (McWatters et al., 2007). How the circadian parameters of periodicity, phase, amplitude and precision are affected is at the heart of these approaches, depicted in principle on Figure 1.2. Circadian time can be defined as the standard of time based on the free-running period of a rhythm where the onset of activity of diurnal organisms defines circadian time zero. Here the ”period” is the time required to complete one rhythm cycle, ”phase” is referred to the given point in the cycle where rhythmic features peaked/troughed. We typically used peak position as the given phase value, the ”amplitude” is defined as the difference between the peak (or trough) and the mean value of a wave and ”precision” defined being the magnitude of the error evident in a lack of robustness in curve fit (Hanano et al., 2006). Future clock models should be able to predict how, for example, how subtle allelic variants would lead to expressed-traits, and effects upon the clock parameters. They could also predict how dynamics at the protein level, taken together with cues like the interactive photoreceptor complex (Devlin and Kay, 2000), feeding information into the system, generate the robust rhythm. It is likely the understanding of natural circuits, synthetic circuits and mathematical modelling jointly will lead tostill further- models and account for all these inputs yielding reproducible results. The research on the plant circadian clock has fascinated many researchers with expertise in numerous domains giving rise to several models. Although the molecular components of the central circadian oscillator in plants have been established, how the circadian clock is organized on a genome-wide scale remains largely unknown. A complete circadian cycle can be fully described by four parameters: period, phase, amplitude, and robustness. How these change due to induced and natural variation is of critical understanding of the internal mechanisms.

The present cornu copiae of expression data both from the laboratories and in silico is being exponentially translated into significant findings. Next generation DNA sequencing technologies are driving increasingly rapid, affordable and high resolution analyses of plant transcriptomes through sequencing of their associated complementary DNA populations. Microarray technologies together with RNA-seq have proven themselves to be a powerful and instrumental tool with a remarkably diverse range of applications. They allow elucidating regulatory gene networks, revealing how plants respond to external cues allowing a better understanding of the relationships between genes and their products, and uniting the ”omics” field of transcriptomics, proteomics, and metabolomics into a now common systems biology paradigm. The core definition of omics is ’the cataloguing of comprehensive sets of biological information from a given sample including genes (genomics), transcripts (transcriptomics), proteins (proteomics), and metabolites (metabolomics) (Chow and Kay, 2013). The application of the omic technologies is especially suited for the analysis of the plant clock because of the multi-loop feedback architecture, the involvement of a large number of genes, time-varying effects, and regulation that occurs at multiple levels. The present insights into the mechanistic connectivity between clock genes and its output processes have greatly broadened through genomic (cDNA libraries, yeast one-hybrid, protein binding microarrays, and ChIPseq), transcriptomic (microarrays, RNA-seq), proteomic (mass spectrometry and chemical libraries), and metabolomic (mass spectrometry) approaches (Chow and Kay, 2013). The efforts of bioinformatics, computational biology and network science are allowing one to delineate the ambiguities of the plant circadian clock. This highly interconnected network is highly manageable by systems approaches as the conducted research will show. These methods are explorative yet joined with emerging technologies like the robust CRISPR/Cas genome editing tools for the introduction of SNPs for example will allow effective translation of findings (Belhaj et al., 2013).

1.2 ACROSS THE BRIDGES AND INTO THE TREES

The network theory which is aiming to quantify biological complexity is indispensable in the study of complex systems (Barabási, 2011, Barabási and Oltvai, 2004). The network takeover is the hallmark of this era. Network science is an attempt to understand networks emerging in nature, technology and society using a unified set of tools and principles. Despite apparent differences, many networks emerge and evolve, driven by a fundamental set of laws and mechanisms, and these are central in the domain of network science. Many networks are driven by a fundamental set of laws as these are the quintessence of this quest.

Networks represent classes of relationships. In a biological context, many different types of node relationships can be measured. Such include physical interactions between proteins or genetic interactions which can be revealed by combinations of mutations. Whereas physical interactions occur between biomolecules in direct contact, the regulatory interactions are directed either activation/inhibition events. For example, in gene expression regulation, a transcription factor is connected to its targets by directed edges. Genetic interactions further connect genes whose parallel genetic perturbation leads to a phenotypic result alternative than one expected from the combination of single effects. For example, synthetic lethal interactions connect genes that weakly affect organism viability when considered individually, yet are lethal when deleted in concurrently (Poyatos, 2012). Similarity relationships link biological objects that are similar according to a common attribute. Many different similarity measures can be used, such as protein sequence similarity and gene coexpression based on correlated transcriptional profiles. Similarity relationships are useful to identify groups of functionally related genes or proteins, the starting building block of a network. Considering that the best protein function prediction algorithms substantially outperform widely used first-generation methods is truly promising (Radivojac et al., 2013). In attempting to understand the entire system the crème de la crème is the integration of different types of interactions with both explorative and validation aims. In a nutshell biological systems when cast in networks allow one to infer the organization of the system in terms of interaction, identification of hubs, inferring the likelihood of disruption, vectorial information flow and information flow through loops, ability to process signals to affect input, and output relationships. In 1873 Eugenio Beltrami described a new approach a special case of Singular Value Decomposition. Through the years to come many more detailed results were published, most notably Karl Pearson's discovery of Principal Component Analysis (Beltrami, 1873). Since that time a series of ideas were translated into improvements of dimensionality reduction techniques as applied to time series data yielding highly interesting results.

”Every object that biology studies is a system of systems,”Francois Jacob. In plant biology the so called mechanistic approaches are no doubt valuable, yet their relevance may diminish as a function of a raised call for understanding of module design. Like Systems Biology beforehand, Synthetic Biology can be viewed as both a tool and a scientific approach for understanding and furthering basic science. Robert Solow has showed that over the past half century, more than half of the growth in the US GDP has been rooted in scientific discoveries, the kinds of fundamental, mission-driven research that we do at the labs (Reikard, 2011). Now is the time that synthetic biology will clearly capture a significant proportion of these findings. Bioinformatics has a unique status vis-à-vis science and technology. The evolution of all these specific domains is of great benefit to the study of complex systems in nature. ’Whereas the insights of Fisher, Turing, and Shannon now underpin the standard repertoire of bioinformatics tools, the theoretical intuitions of Delbrück and Gamow drive those tools, and the empirical sensibilities of Sturtevant, McClintock, and Nirenberg are embedded in them’ as David Searls once stated (Searls, 2009). In terms of the circadian network the groundbreaking work of Steve Kay, Andrew Millar, Barabasi, Altschul, Ashburner and Altman will surely result in significant crop biotechnology applications and will make such a network instrumental, leading to potential and realistic manipulations at every level.

Plant genomes are very distinct and difficult to comprehend as they can be by far larger, even 100 times larger than some mammalian genomes. They are characterized by much higher ploidy, which is estimated to occur in up to 80% of all plant species, and higher rates of heterozygosity and repeats. Furthermore, the gene content in plants can be very complex, as result of genome duplication events. These and other characteristics make them a special case from the perspective of bioinformatics and sequencing. Many of the agronomically important traits, such as yield and drought tolerance, involve multiple genes and complex interactions with the environment calling for sophisticated pipelines. Bottom up approaches, painstakingly produce models and great emphasis should be placed on these. Whereas top down approaches with the present capabilities should clearly be incorporated, it is not one instead of the other. It seems the present mission of systems biology will be the fusion of both. We are on the edge of another shift, from producing ontologies reliant upon individual experiments we are learning from big data giving rise to ’network learning’ as discussed by Dutkowski (Dutkowski et al., 2013). This Thesis sought to delineate the central circadian actors, transcription factors and their interactions over time. The hypothesis being that knowing those central TF will allow one to infer the dominate phase modules and lay grounds for their interchange. Thorough understanding forms the foundation of modification of such networks. In this panorama, not only from a scientific perspective but also for commercial or strategic motivations, the identification of the principal actors inside a network or inside a community is truly important. The interpretation of these data opens up new fascinating research questions which will be discussed in detail.

1.3 ORGANIZATION

The main contributions of this thesis, therefore, can be summarized as follows:

- Chapter 2 introduces some fundamental concepts and methodologies that will be widely adopted, the formal conventions used and the implemented algorithms that will be discussed throughout the Thesis. First, we define some formal mathematics conventions used to define the terminology, notations and a few other notations typical of the network theory. Moreover, we formalize the characteristics of a graph and its properties. The metrics used together with entire workflows are introduced.

- Chapter 3 is intended to answer the most fundamental question in circadian clock biology: what are the genes moderating circadian rhythms, both the core genes and their targets? It defines the circadiome, the extent of circadian clock control in the model plant. Chapter 3 introduces the dimensionality reduction used, the *cis* 'elementome' and machine learning methods. Publicly-available transcriptomic data were obtained from a number of circadian time-courses. The aim was to identify a small number of key patterns (components) in the data which could best represent the dataset as a whole (dimensionality reduction). It was assumed that as no environmental variables were present in the experiments, identified patterns ought to represent internal variations (circadian rhythmicity). Independent Component Analysis was chosen as the method of dimensionality reduction. The *cis* elements contributing to more than one phase of expression or forming part of previously recognised light or circadian elements are shown for each of the four phase modules based on projection onto circadian ICA components and discussed in Chapter 4.

- Chapter 4 introduces the reader to the Random Forest performed on single and interacting features. More-specifically, random forest is an ensemble machine learning method in that it combines the results from multiple decision tree classifications to obtain even higher performance. The tuning and extended application will be discussed from the perspective of different input features. The genome-wide analysis of the *cis*-elements in circadian promoters should identify the motifs that control rhythmic RNA expression of a clock-controlled gene, and this facilitates the identification of the trans factors that create such rhythms. Translating this into machine learning approach given a gene the aim is delineate the features that ideally group it within a particular module. Essentially the aim is the programming of gene expression with these combinatorial promoters

- In Chapter 5 the properties of the circadian static network are discussed. The central focus is the episteme and the techne of the plant circadian modules. It is divided into 3 subparts with the first one being the static properties of the circadian system. Next the focus is on the identification of plant modules, specifically the community detection methods and the proxy measures. The chapter concludes with linking phenotypic and molecular networks across multiple types of circadian datasets from the model plant. Static interactions are compared to differential interactions and specific quantitative proofs of the findings regarding the communities are presented (Bandyopadhyay et al., 2010). Fundamentally the properties of the differential circadian system are presented across two different experimental designs. That brings a lot of novelty to the computational and circadian components of the current state of circadian knowledge and computational biology. The comparative genomics of circadian clocks using novel comparison methods are discussed. Having the potential applicability to one or more of the following mentioned crops like maize, rice, wheat, sorghum, cassava, soy to name a few is of great agronomic importance. The results of the performed experimentation are presented, showing that our approach is able to generate reproducible results. Concluding, we discuss a possible application of this methods to devise a novel conservation detection tool. Over the past years, our knowledge of the plant circadian system has mainly emerged from studies in *Arabidopsis thaliana*. Model systems are providing an indispensable platform to understand clock architecture and circadian-regulated development and physiology. The analysis of circadian clocks in the green lineage gives insight into evolutionary processes in diverse species of plants. This information is being exploited to study the relevance of circadian regulation to crop performance first on gene and then on network level.

- In Chapter 6 the regulatory network on circadian transcription factors is presented. Both the existing and novel methodologies are discussed and the chapter focuses on specific transcription factor target examples. Transcription factors play a central role in the regulation of gene expression. Their interaction with specific elements in the DNA mediates dynamic changes in transcriptional activity. These are central to the understanding of the plant clock and hence particular attention is placed on a series of them acting alone and in combination with other transcription factors. This provides a comprehensive picture of the regulatory and mechanistic pathways contributing to circadian function providing first such circadian regulatory network.

- In Chapter 7 the potent applications to the synthetic oscillators and control systems are discussed (Liu et al., 2011, Nepusz et al., 2012, Slusarczyk et al., 2012). Topological vulnerability of the network is discussed, one learned in an unsupervised manner. Such approaches are compared to the experimental results of loss of function of the core clock genes, which may result in a short period clock and arrhythmicity. The application of higher order methodologies, tensors and other decompositions are proposed. Chapter 7 concludes with the review of present findings and future directions with focus on findings from induced and natural phenotypes. As the challenge of global food security has brought plant research and breeding into sharp focus, having further directions mapped out is critical for more experiments to be pursued. Progressive promoter element combinations have been detected and these classify conserved orthogonal plant circadian gene expression modules. Differential interactions provide insights across space and time and decomposition methods allow for cross examination of multiple types of data. Results depicted here raise further questions that have to be answered and which will be, through careful monitoring of the oscillator and its natural and induced phenotypes. Plant biotechnology has enabled improved farming techniques and crop production around the globe by increasing plants' resistance to diseases and pests, reducing pesticide applications and maintaining and improving crop yields. This chapter offers the links between the clock, its outputs and such potential applications. "A major challenge in systems biology is to create quantitative, predictive models of gene-circuit dynamics" Uri Alon once stated (Alon, 2003). The present work is certainly showing us that we have the episteme and the techne and the integration of information can indeed lead to the in depth understanding of the plant clock.

We might be hitting the limits of reductionism yet this thesis will remain a mixture: an interplay between complex network analyses and reductionist views upon individual circadian genes. Each section that describes novel contributions is placed in context of the present state of knowledge including the original publications where the results first appeared. The main contributions of this thesis, therefore, are presented in the chapters to follow.

'Intellectual inquiry is worthwhile in itself'.

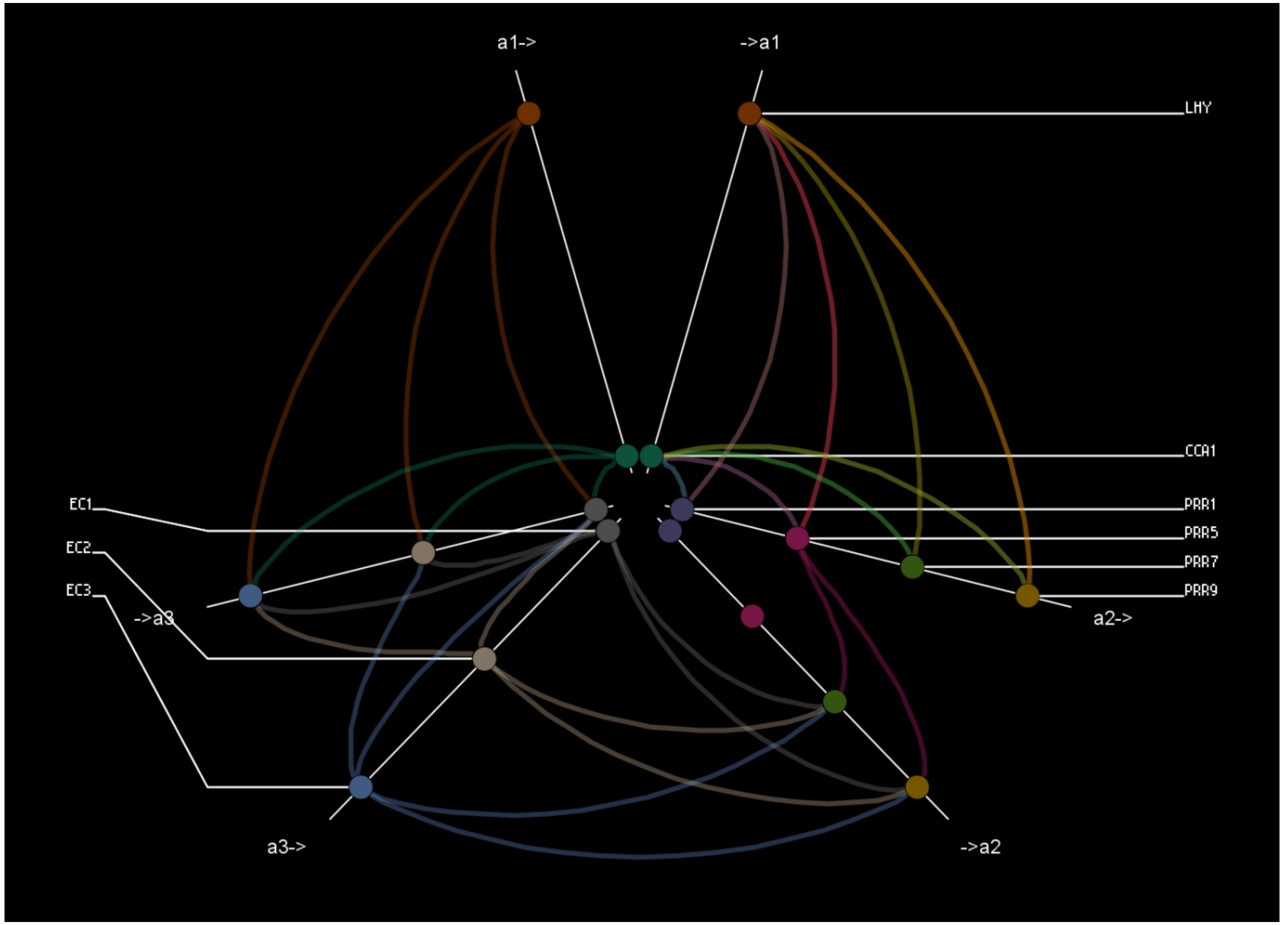


Figure 1.1: The Core of the Plant Circadian Clock Plot. The colours represent individual transcription factors. The transcriptional view of the clock can be summarized as 3 loops: the core loop, the morning and the evening loops. The core loop of the plant circadian clock is comprised of three inhibitory steps: *LHY* and *CCA1* are expressed simultaneously at dawn and act to repress expression of *TOC1* at the same time activating that of *PRR5*, *PRR7* and *PRR9*. Expression of *LHY* and *CCA1* is subsequently inhibited during the day through the action of *PRR9*, *PRR7*, *PRR5* and *PRR1/TOC1*. Expression of the PRRs is then shut down during the night through the action of the *ELF3*, *ELF4* and *LUX* proteins being component of the Evening Complex, EC. This allows *LHY/CCA1* transcription to rise again at the next morning. Moreover, *LHY/CCA1* act at the same time negatively on the components of the EC. Repressive interactions from the latter components of the PRR cascade onto the former ones may enable their temporal separation and expression as consecutive modules. The EC also feeds back onto its own expression. The axes represent the 3 modules whereas their separation illustrates the interactions of the components within them

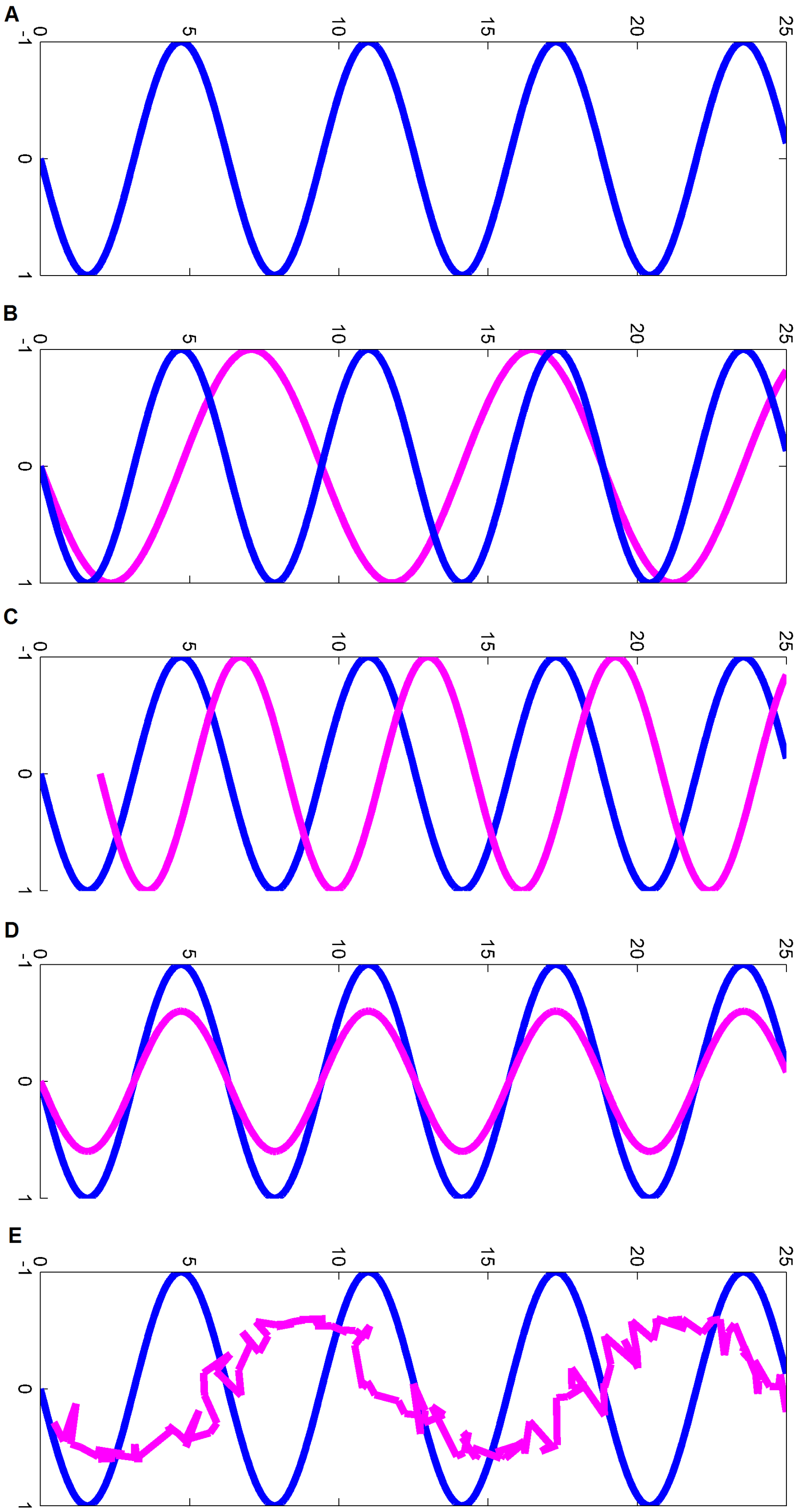


Figure 1.2: The Alterations in the Circadian Rhythm. The phase can be defined as the relative angular displacement between a periodic quantity and a reference angle and the period can be defined as the time elapsed for one complete oscillation, cycle. Panels A through E portray the changes in the parameters given internal and external stimuli. Subsection A portrays a representative circadian rhythm. Subsection B represents change in period. Subsection C represent delayed onset, resetting of the clock. Subsection D represents reduced amplitude of oscillation whereas subsection E represents aberrant rhythm

Chapter 2

METHODOLOGIES

FORMAL CONVENTIONS

Chapter 2 introduces some fundamental concepts that will be widely adopted throughout the thesis. The chapter explains the major algorithms and methodologies used and these are supported by the pseudocode (Appendix).

2.1 CIRCADIAN NETWORK TERMS AND CONVENTIONS

A network is defined as an object composed of entities (here genes) interacting with each other through certain type of connections (Pearson correlation, mutual information). The natural means to mathematically represent a network is through a graph model. Here we define a graph $G = (V, E)$ as the abstract representation of a set of V, nodes (vertices) and a set E, *edges* that connect vertices together. Thus, we denote as $V(G)$ such set of vertices V and, similarly, as $E(G)$ the set of edges within the graph G. Two vertices connected by an edge e are called endnotes for e, and they are adjacent. Hence the adjacency matrix, denoted by X contains entries which indicate whether two vertices are adjacent. Usually 1 denotes edge presence whereas 0 denotes edge absence. As the graph is undirected, the matrix is symmetric with respect to its diagonal. A gene-coexpression network can be informally defined as a network representing the interactions among genes using a preselected measure. Specifically a gene-coexpression network represents the behaviour of a set of genes, that act in response to a given stimulus. These data are obtained via for example microarray experiments in which expression is sampled for example across time using a designated number of replicates.

That allows one to work with a matrix which is further used to calculate the pairwise correlation between a pair of nodes for example via Pearson correlation.

2.2 THE APPLIED ALGORITHMS AND METHODOLOGIES

The upcoming section focuses on all the fundamental tools that were used in throughout the thesis. The thesis aims to find the initial points of convergence between bottom-up and top-down models which will provide a thorough understanding of the links between the intrinsic mechanisms of the clock and the phenotypic responses that are circadian regulated. For these links to be made a series of techniques were applied to the data to capture the meaningful signal representing true associations within the data. Unless noted otherwise, we controlled the False Discovery Rate for all analyses using the Benjamini and Hochberg procedure.

2.2.1 DATASETS

The following transcriptomic datasets were used in this study: 1. GSE8365 Arabidopsis thaliana, Affymetrix Arabidopsis ATH1 Genome Array (Covington et al., 2008) 2. GSE5612 Arabidopsis thaliana, Affymetrix Arabidopsis ATH1 Genome Array (Edwards et al., 2010) 3. GSE23918 Zea mays 105K Agilent Microarray (Hayes et al., 2010) 4. GSE31763 Zea mays Affymetrix Maize Genome Array (Khan, 2010) 5. GSE28124 Oryza sativa 57K Affymetrix Rice Whole Genome Array (Xu et al., 2011). 6. used experimentally E-MEXP-1304 (Michael et al., 2008)

Microarrays were used to identify circadian regulated genes of Arabidopsis thaliana with the aim being delineating key circadian modules and hence later on downstream elements. To do so it is important to, firstly, understand how the clock regulates outputs and, secondly, determine which pathways and processes may be under circadian control. With regards to the first study, the researchers have grown the groups of Arabidopsis seedlings in light/dark cycles for seven days. Then these were transferred to constant light. After 24 hours in constant light 12 samples were harvested at 4 hour intervals over 44 hours, for RNA extraction and hybridization on Affymetrix microarrays.

2.2.2 PREPROCESSING OF THE DATA

The pre-processing procedure (Holter et al., 2001) involved log_2 transformation; centring of the columns by subtracting the average; column normalisation; centring the

rows by subtracting average; then row normalisation. The preprocessing procedure involves (Holter):
Step 1: log2 transform
Step 2: centre the columns by subtracting the average
Step 3: column normalization
Step 4: center the rows by substracting average
Step 5 row normalization
From the resultant data, covariation matrix was constructed and then ICA was performed.
The resulting gene's transcriptional responses had mean of 0 and unit standard deviation.

2.2.3 PRNCIPAL COMPONENT ANALYSIS

The microarray gene expression data are traditionally represented as a $n \times m$ matrix X with n genes under m samples or conditions. The columns of the dataset denote gene expression signatures of n genes during m environmental conditions while each row of matrix X represent gene expression profiles of each gene across all m conditions. PCA is a linear transformation of data from genes × array space to eigengenes × eigenarrays space. The microarray data can be considered a cloud of n points in a d dimensional space. Let matrix E of size N_g(genes) × N_t(arrays) tabulate expression data. Element $E_{g,t}$ is the relative expression level of gene g at time t. Gene g is represented by N_t numbers, that is the components of a vector in N_t dimensional space. Genes form a cloud of N_g points in N_t space where the number of samples is less than genes. In order to find directions of greatest variance covariance matrix was computed and eigenvectors of covariance matrix with the greatest eigenvalues were assessed. Essentially, matrix E was decomposed into the product $E = U\Sigma V^t$ of a $N_t \times L$ orthogonal matrix U. a diagonal matrix Σ and $N_g \times L$ orthogonal matrix V where $L = rank(X)$. Here the columns of U are eigengenes and the columns of V are eigenarrays where both are uncorrelated. The eigengenes and eigenarrays are unique except for phase factor of ± 1, such that each eigengene captures both parallel and antiparallel expression patterns. The fraction of eigenexpression gives the relative significance of the given eigengene and eigenarray in terms of the overall expression that they capture. PCA was applied to all 6 datasets and the results were compared. Of interest is the linear combination of eigengenes. The remaining coordinates represent the common variations in transcriptome exposed to certain conditions. As the mRNA levels constitute one of many variables contributing to the dynamics of the system it is difficult to envision how the hyperplane is located in the system model space, that is to what extent these profiles recapitulate the circadian system. The linearity assumption is a crude approximation of the biological system. Yet linear models can show dominant trends through inspection of values across time relying on linear effects on mRNA levels.

2.2.4 INDEPENDENT COMPONENT ANALYSIS

ICA measures the relevance of a linear combination $a^T x$ in terms of the size of its absolute kurtosis. Fundamentally PCA allows detection of L uncorrelated profiles whereas ICA allows for detection of L independent profiles. Lack of correlation determines the second-degree cross-moments of a multivariate distribution and statistical independence determines all cross moments. Independence is a high-order statistic that is a much stronger condition than orthogonality. The FastICA package for Matlab (http://research.ics.aalto.fi/ica/fastica/) was used to carry out ICA on the microarray dataset. As the FastICA algorithm relies on random initialisations for its maximization and faces the problem of convergence to local optima, we iterated FastICA 100 times and took the average in order to alleviate the instability of the slightly different results in each iteration. ICA measures the relevance of a linear combination $a^T x$ in terms of the size of its absolute kurtosis. After pre-processing and normalisation, the ICA model for gene expression data can be expressed as: $X = AS$. In this ICA model, the columns of $A = [a_1^T, a_2^T, \ldots, a_m^T]$ are the $n \times m$ latent vectors of the gene microarray data. Each column of A is associated with a specific gene expression mode. S contains the $m \times m$ gene signatures where the rows of S are statistically independent to each other. The gene profiles in X are considered to be a linear mixture of statistically independent components S combined by an unknown mixing matrix A. Once latent variable matrix A was obtained, the corresponding elementary modes were identified to extract information for classification.

2.2.5 EMPIRICAL SIGNIFICANCE TEST FOR ICA

To test whether d independent components are significant in representing the whole microarray data, we designed an empirical significance test. The idea is that, if the data matrix X can be properly represented by d independent components, then the reconstruction error $re = ||X_{n,m} - \hat{A}_{n,d}\hat{A}_{d,m}$ should be small, where $\hat{A}_{n,d}$ and $\hat{A}_{d,m}$ are estimated by an ICA algorithm, and $||||$ denote the Euclidean distance between X and the reconstruction $\hat{A}_{n,d}\hat{A}_{d,m}$. As a result, randomised reconstruction errors rre, which are obtained by randomising X, learning A and S, and calculating the reconstruction error, have a probability to be larger than re. Based on this idea, we report an empirical p-value to test how much d independent components are significant in the ICA analysis by the following procedure. For the original data X, we ran ICA, and obtained re. We then randomized X to RX by a rotations-based procedure SwapDiscretized (Ojala et al., 2009), which guarantees that the distributions of the discretized values in the rows and columns do not change. For RX, we ran the ICA procedure to obtain a randomized rre. We then repeated the procedure 1000 times so that we had 1000 randomized $rres$. Finally, the empirical p-value was the frequency that $rres's$ are smaller than $re's$.

2.2.6 SIGINIFICANCE OF EIGENVALUES AND EIGENGENES

The equal importance of profiles within *Arabidopsis* was assessed using a comparative mathematical framework as proposed by (Alter et al., 2003). The antisymmetric angular distance between the datasets was used to assess the equal significance of eigengenes with regards to ratio of expression captured by the first 2 eigengenes. Angular distance of 0 indicates significance in the second data set given the first data set and no significance being characterized by a distance of $\pm\pi/4$. $E_1 and E_2$ denote eigengenes from first and second dataset respectively, whereas N denotes consecutive eigengene.

$$\theta_m = \arctan(E_{1,N}/E_{2,N}) - \pi/4 \tag{2.1}$$

The values were almost 0 indicative of similar importance ($\measuredangle 4, \measuredangle 2$).

2.2.7 PROJECTION DOWN TO A SUBSPACE AND GOODNESS OF EMBEDDING

Projection of data into ICA subspace can reveal structures in the data that may be used for classification. Projection scatter plot coordinates $q_{i,k}$ for transcriptional response n_i projected on component v_k are calculated as $q_{i,k} = n_i v_k$. Projected were the gene transcriptional responses onto the circadian components. A stringent radial cutoff of 0.8 was used, a cutoff that holds for circadian genes across all *Arabidopsis thaliana* datasets. The distance of each gene from the origin is its amplitude of expression in the subspace spanned by the circadian eigengenes. One may expect that genes that have all their expression in this subspace close to $r{\approx}1$ are circadian regulated and those close to $r{\approx}0$ are not as confirmed by the known set of circadian genes. The genes are later sorted with regards to their angular distance.

2.2.8 MODULE DETECTION AND PRESERVATION

We define a network as $G = V, E$, where V is the set of nodes and E is the set of edges. Alternatively, we represent a network by its adjacency matrix, $W = (w_{ij})$, where $w_{ij} = 1$ if there is an edge between v_i and v_j, and 0 otherwise. Let s_{ij} be the similarity between gene i and gene j, where similarity in this study is measured by Pearson correlation coefficient and (MIC).

Two types of analysis were performed. One involved assessment of similarity of eigengenes across species, the second involved assessment of conservation of circadian orthologous in terms of transcriptomic profiles. Comparative mathematical framework was used for the initial purpose. The framework involved linear transformation of the two data sets. Angular distance was a measure of significance. The hand in hand correspondence is the foundation for the workflow.

Integrative correlation coefficient was used for that purpose as a measure of cross study reproducibility for gene expression array data. We have two microarray studies S_a and S_b with sample size of n_a and n_b and a total of m common genes. All genes conform to the mean zero and variance one. The expression vector for a gene x in study S_a is notated as x_a. A denotes the $(m-1) \times n_a$ data matrix for study S_a without gene x. Similarly B denotes the $(m-1) \times n_b$ data matrix for study Sb without gene y.

Furthermore cE_m is the $m \times m$ matrix with every element equal to c and I_m denotes the $m \times m$ identity matrix. If x and y are two random vectors of length m, then $[I_m - \frac{1}{m}E_m]x$ returns $x - \bar{x}$, $[I_m - \frac{1}{m}E_m]y$ returns $y - \bar{y}$, and $cov(x, y) = y^t[I_m - \frac{1}{m}E_m]^2x$. Denote $J_m = [I_m - \frac{1}{m}E_m]^2$. The integrative correlation coefficient for gene x in studies S_a and y in S_b can be defined as

$$\frac{x_aA^tJ_{m-1}Bx_b^t}{\sqrt{(x_aA^tJ_{m-1}Ax_a^t)}\sqrt{x_bB^tJ_{m-1}Bx_b^t}} \tag{2.2}$$

2.2.9 JACCARD INDEX FOR MODULES

We used the Jaccard similarity score $TruePositive/(TruePositive+FalsePositive+FalseNegative)$ to asses module composition. The similarity of two modules is measured by the Jaccard index score between the edges of two co-expression graphs whose nodes are the members of the modules and whose edges are those pairs with a co-expression $>= 0.8$. This involves permutation to derive significance scores.

2.2.10 CLASSIFITCATION OF GENES GIVEN THEIR CIS ELEMENTS

Knowing the classes of expression patterns it is of interest to find the motifs. Motifs are representations of a binding site of a TF on its target. These are represented by stings standalone, strings with mismatches and position weight matrices. Can I leave this here as it is important I tried many yet not so for the biological restuls. A number of motif discovery tools were developed over the last decade including Gibbs Sampler, MEME, Weeder, Improbizer, DME, DEME, A-GLAM, RED2. They differ by the learning principle employed to infer the model parameters and by their capability of learning the position distribution of the binding sites. For computational approaches, the fundamental improvements include searching for differentially abundant motifs and learning a position distribution. DISPOM uses discriminative learning principle and the position of a motif from the start site can be learned from the data and hence was used in this study. It includes a heuristic for finding motifs of unknown lengths. Motifs were discovered by searching the promoter sequences (1000 bp upstream region) of genes found to be circadian, one motif at a time. These are the motifs to be sebsequently used in the phase group classification. Two distinct matrices were created. X number of motifs was found to be overrepresented on a gene set with a FDR less than X. To avoid redundancy the motifs were clustered using a hierarchical agglomerative clustering algorithm. A library of published motifs of known transcription factors was screened (JASPAR) This was a important step in verification of the relevance and understanding of the motif groups as many cis elements were described previously. The information about motifs in promoters can be represented in the form of a matrix M where each element M_{il} annotates the number of occurrences of motif m_l in the promoter of target gene n_i. Matrices were constructed relying on overrepresentation analysis relying on all combinations of 8mers of letters A,T,C,G. Altogether 48 features were preselected. Classification was performed using Random Forest for MATLAB (http://code.google.com/p/randomforest-matlab/). To measure performance 2/3 cross validation was used. The performance was assessed using receiver operating characteristic (ROC) curves with true positive (TP) being gene correctly classified as circadian regulated and true negative (TN) being correctly classified as not circadian. ROC curve is a plot of sensitivity versus (1-specificity). It is created by changing the threshold one places in the majority vote of decision trees. In addition area under the receiver operating characteristic curve (AUC) was calculated. The motif positional information and sequence content is both the output.

Predictions were evaluated using difference in probabilities. The probability difference is the difference in prediction of the model having the knowledge about the value of A_i representing a feature, when it is absent.

$$probDiff_i(y|x) = p(y|x)_i p(y|x\ A_i) \tag{2.3}$$

2.2.11 ASSESSMENT OF PERFORMANCE

ROC/CROC (Swamidass et al., 2010) curves were used for assessment of the obtained predictions. Figure 2.1 illustrates how each ROC curve is derived in principle. The figure describes the assessment of performance. It is important but where should it be?

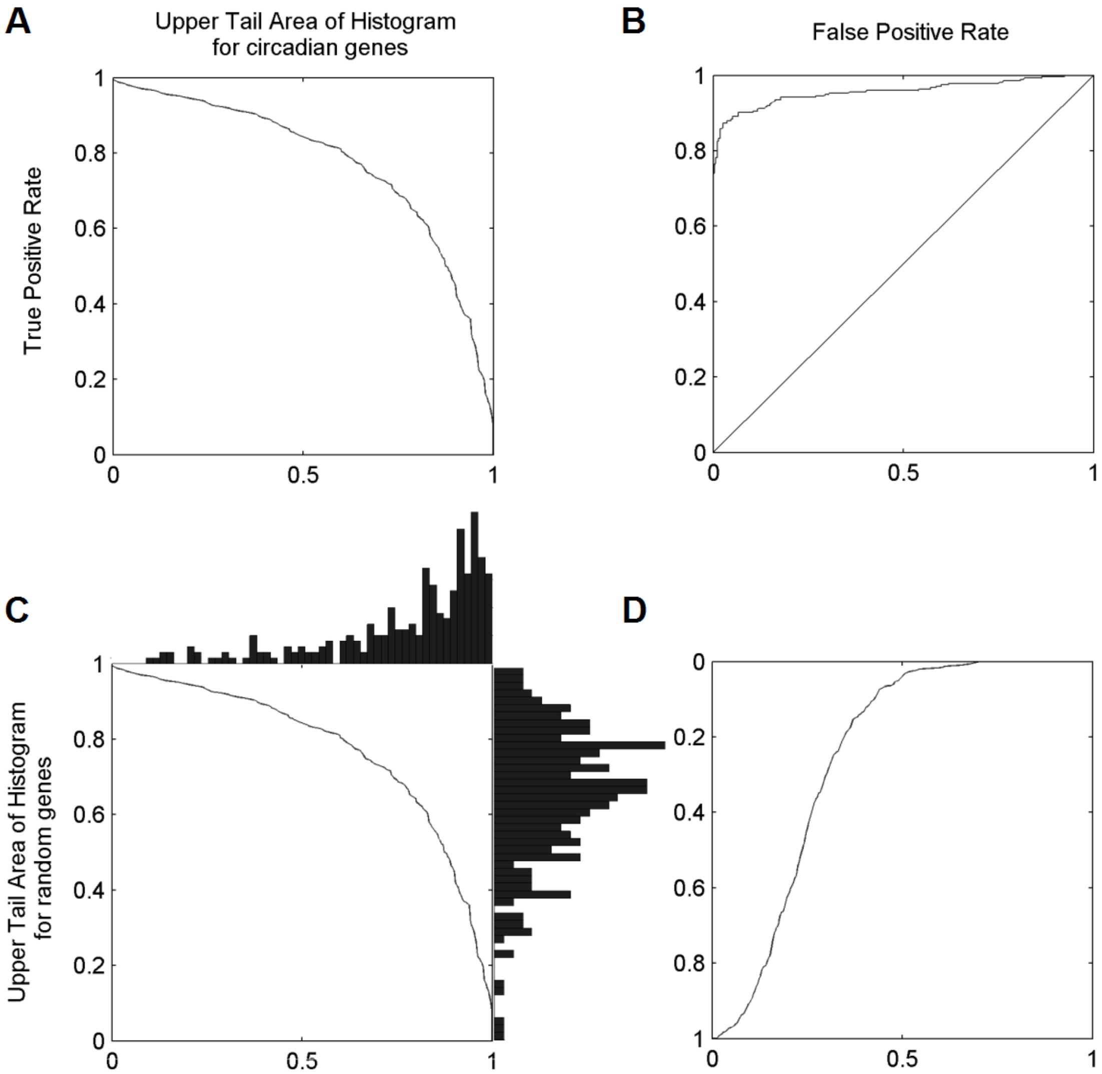


FIGURE 2.1: For each classification analysis the TPR can be read of from A and FPR can be read of D. The pair (x, y) is plotted as B. This is shown from the perspective of cutoff points in C. One can trace the formation from C through A and D to B.

2.2.12 DYCK PATH ANALYSIS

The intention was to carry out PCA on trandformed data tree objects. The adapted dyck path code is included in the Appendix. Method involved several steps. First step is to create adjacency matrix based on each random forest tree generated. Each edge from each random forest tree is annotated in adjacency matrix with value of one. Second step is to vectorize all of these random forest trees to a vector and subsenseqently combine them into a matrix. Matrix got normalized and the PCA was performed on the tree-matrix. The matrix with the code is deposited in the appendix. Based on the PCA results cis motif interactions with the highest loadings were selected.

2.2.13 THE LARGE ARABIDOPSIS DATA

All present up to date experiments were collected and concatenated into one array of data for the purpose of verification of results (Appendix). This was used at the ground truth, gold standard at several stages of this investigation.

2.2.14 ALTERNATIVE MODULE DETECTION USING CLUSTERONE AND TOPOLOGICAL OVERLAP MEASURE

The steps of constructing a coexpression network and module detection resembled concepts proposed in Dong (Dong and Horvath, 2007). The concordance of gene expression was measured with Pearson correlation. Such matrix was dichotomized to arrive at adjacency matrix for the unweighted network. The topological overlap dissimilarity was used as input of the hierarchical clustering so essentially modules were branches of a hierarchical clustering tree. Modules were approximately factorizable in accordance with the proposition in Dong (Dong and Horvath, 2007). It is important to characterize gene expression data X that lead to an approximately factorizable correlation matrix. That was adhered to in accordance to the proposition below:
The adjacency matrix A is approximately factorizable if there exists a vector CF with non-negative elements such that

$$a_{ij} \approx CF_i CF_j foralli \neq j \tag{2.4}$$

for all $i \neq j$ CF_i is referred to as conformity of the $i-th$ node (Dong and Horvath, 2007).
The aforementioned topological overlap dissimilarity which is used as input for hierarchical clustering is defined as the equation below with k_i and k_j being number of direct

neighbours :

$$TOM_{ij} = \frac{\sum_u a_{iu}a_{uj} + a_{ij}}{min(k_i, k_j) + 1 - a_{ij}} \tag{2.5}$$

$$DistTOM_{ij} = 1 - TOM_{ij} \tag{2.6}$$

Clustering with Overlapping Neighbourhood Expansion known as ClusterONE was implementated and is available at http://www.paccanarolab.org/cluster-one/. Starting from a single seed vertex, a greedy procedure adds or removes vertices to find groups with high cohesiveness. This is repeated from different seeds allowing overlapping groupings. The second step is the quantification and the merging of overlaps above a threshold. Next complex candidates are discarded in accordance with the parameters. ClusterONE was excessively tested on a large scale data set and compared to a series of approaches (Markov cluster, molecular complex detection, affinity propagation, restricted neighbourhood search clustering algorithm, CFinder, clustering based on maximal cliques and repeated random walks) (Nepusz et al., 2012).

2.2.15 CONSTRUCTION OF REGULATORY NETWORKS

The objective is recovering regulatory networks from gene expression data. The obtained networks are directed graphs with p nodes, where each node represents a gene, and an edge directed from one gene i to another gene j indicates that gene i (directly) regulates the expression of gene j. The connection may indicate either an activator or a repressor. The notion of dependence is instrumental to model gene regulatory networks. Some of the methods used to identify dependence between random variables include Pearson's, Spearman's and Kendall's correlations; distance correlation; mutual information and maximal information coefficient. They are used to identify linear associations, exponential associations, quadratic association and local correlation to name a few. In the present analysis several measures were tested these being Pearson's correlation, mutual information and maximal information coefficient. After several tests the selected method was time shift methodology with the pseudocode present in the Appendix. The second method of choice was GENIE3 MATLAB code downloaded from http://homepages.inf.ed.ac.uk/vhuynht/software.html (Huynh-Thu et al., 2010). It decomposes the prediction of a gene regulatory network between genes into different regression problems. Per each such regression problem, the expression pattern of one of the target gene is predicted from the pattern of expression patterns of the remaining genes. This involves Random Forests. In principle 'per gene j, a learning sample is generated with expression levels of j as output values and expression levels of all other genes as input values. A function is learned from the learning sample and a local ranking of all genes except j is computed. The p local rankings are then aggregated to get a global ranking of all regulatory links' (Huynh-Thu et al., 2010). The settings were 1000 trees and default mtry. The results from all the tested methods were assessed. The additional files in appendix include the TF and the target list.

2.2.16 INFERRED DIFFERRENTAL NETWORK CODE

The differential network analysis operated on the idea that the differential interaction score is computed for each gene pair, by subtracting the static score in the first condition from the static score in the second, then indexing this value against the null distribution of values (by comparing 2 consecutive days). Interaction in which the obtained score is higher than that of condition 2 is positive differential and interaction in which condition 1 is less than that of condition 2 is negative differential. This methodology was adapted to meet the biological hypotheses. Gene expression measurements were divided into two categories: day and night (across 2 consecutive days). The 2 equally sized (day and night) correlation matrices on identical set of circadian genes were constructed. The day and night specific interactions (edges) between genes (nodes) are selected by differential mapping. The pattern of associations for each set was investigated by construction of individual networks. The Pearson correlation was considered as a measure of the underlying interdependencies between variables. To place the results on a statistical footing we performed a two sample Mann Whitney test. To define which edges are included in the differential networks we set a cut-off on the two-tailed p-value. For the same sample size, the power to estimate correlations is lower than that to estimate a change in a single variable. Therefore, we choose to set the cut-off for the inference of the differential network based on the uncorrected threshold being $p < 0.01$. An edge in the differential network is included if the correlations between two nodes is significantly different across the two groups. The association network inferred under this set up is the corrected differential network.

2.2.17 SIMPLE NETWORK STATISTICS

Networks were created using cytoscape and its plugins and using GEPHI, iGRAPH and nodetrix. The adjacency matrix, denoted by X contains entries which indicate whether two vertices are adjacent or not. This matrix X of size $n \times n$ can well describe an unweighted undirected graph $G = (V, E)$ containing n vertices. Rows and columns of the adjacency both represent the index of each vertex in the graph, and are labeled as $1, 2, \ldots n$. Each entry x_{ij} in the adjacency represents if the indicated pair of vertices n_i and n_j are adjacent or not. Usually, there is a 1 in the (i, j)th cell if there is an edge connecting n_i and n_j in the graph, or a 0 otherwise. Thus, if vertices n_i and n_j

are adjacent $x_{ij} = 1$, otherwise $x_{ij} = 0$. Because the graph is undirected, the matrix is symmetric respect to its diagonal.

We computed certain network statistics to confirm that our network is not a randomly generated network and has the properties desired in a biological network. The total number of edges present in the network was X which is around 1% of the total possible directed edges in the network, which is an indication of sparseness, common characteristics of biological networks. There is one connected component in the network indicating strong connectivity. The mean shortest path length is X which means that most genes are close to each other and the network diameter representing the maximum distance between two connected nodes is X. These characteristics have been described as small world properties of real networks. Barabasi and Albert used the node degree distribution to distinguish between the topologies of random and scale-free networks. This network shows a power-law like distribution on log scale as shown in the result chapters.

- The degree, k_i, is the network measure that indicates the number of connections that a node i has with the other nodes in the network.
- The degree distribution $P(k)$ represents the probability that a node i has k links. $P(k)$ can be calculated by computing the total number of nodes with degree $k = 1, 2, \ldots$ and dividing by the total number of nodes N.
- The betweenness centrality, b_i, is the measure that indicates how central a node v is in the network and can be calculated by $b(v) = \frac{\sigma(v)}{\sigma}$ where σ denotes the total number of pairs of nodes and $\sigma(v)$ denotes the number of shortest paths that pass through v.
- The clustering coefficient, c_i, indicates to what extent the neighbors of a selected node i are connected to each other and can be obtained for each node i by:

  $$c_i = \frac{2K_i}{k_i(k_i - 1)} \tag{2.7}$$

  where K_i denotes the links observed among the neighbors of node i and k_i represents its degree. For $k_i < 2$, c_i, is defined to be zero.

In this Chapter the formal conventions that will be used in the rest of the thesis and those fundamental concepts related to the gene expression analysis were introduced. That is the framework for the formalization of problems related to network analysis. These concepts will be extensively adopted in the upcoming chapters and many of the experiments are reliant upon them.

Chapter 3

DETECTION OF CONSERVED ORTHOGONAL PLANT CIRCADIAN GENE EXPRESSION MODULES

3.1 INTRODUCTION

Time-series expression experiments are an increasingly popular method for studying a wide range of biological systems. Here, the microarray gene expression timeseries served as the starting point of the analysis. We tested the hypothesis that progressive combinations of multiple promoter elements act in concert and may be responsible for the full range of phases observed in plant circadian output genes.

The identification of circadian genes is, of course, the critical first step for in-depth understanding of the network topology of clock regulation. Much microarray data is publicly available from plant circadian time courses and a range of approaches have been used to identify circadian genes. Identification of circadian genes varies greatly from one method to another and no defined subset of plant circadian genes has been agreed upon (Doherty and Kay, 2010). Existing approaches commonly involve supervised selection of genes fitting to certain predefined patterns. However, such approaches are, by definition, biased and, thus, may not give the full picture of a circadian system. Biological time series do deviate from hypothetically imposed models. Despite the undisputed utility of Fourier Theory for analysis of the frequency domain of time series there are established limitations for short and noisy time series with background trends, multiple frequencies and not harmonic oscillations (detrending, aliasing and Gibbs phenomenon pose further limitations). In most other fields, unsupervised methods are preferred so as not to merely select for features of the system which are already known. Unsupervised methods for gene expression analysis can be divided into two dominant categories: clustering approaches and projection methods. The latter seeks the inherent structure in the data in unsupervised manner. In contrast to clustering, linear projection models permit the inference of combinatorial control where expression levels are described in terms of linear functions of a smaller number of common hidden variables. Linear projection models could, thus, be said to describe a smaller number of regulatory effects. Since genes do, indeed, respond to combinations of only a relatively small number of input variables, the patterns observed via linear projection approaches, perhaps, better approximate real regulatory functions. Furthermore, dimensionality reduction inherent in projection allows for identification of new problem-specific parameters (module characteristics), that are distinctive of the circadian system in plants which would not have been identified by supervised methods of course given the inherent assumption that underlying trends in constant conditions must be circadian. It is worth noting that the existing methods for identification of circadian genes provide a small overlap in terms of identifying the same genes as circadian when tested on the same datasets showing the likelihood of false positives for the liberal methods and false negatives on for the conservative side (Appendix). No benchmark sets have been devised as the set and its definite characteristics are unknown and hence the problem of identification of a reliable set that can serve as foundation for the studies of the system remains (Doherty and Kay, 2010).

With a view to better infer the network topology of circadian output genes, we applied the linear projection method, independent component analysis (ICA), as non-biased method of identifying trends underlying short timecourse circadian microarray data. Historically, the special case of singular value decomposition was described in 1873 by Eugenio Beltrami. The adaptations of the method were used throughout the coming years with the special case of Principal Component Analysis in 1901 for which the proof of existence was presented in 1936 by Carl Eckart and Gale Young (Beltrami, 1873, Eckart and Young, 1936). Here we use several adaptations of the groundbreaking methods for the analysis of the circadian system. Two objective circadian patterns were captured as the most significant trends underlying gene expression profiles in *Arabidopsis*, *Zea mays* and *Oryza sativa*, new characteristics of the system as a whole, which would not have been identified by a direct focus on individual genes. We found that the trends from the analysis reflect conserved structure in the data in that the same eigenvectors were present in all related microarray experiments and were conserved across species. Significantly, these circadian trends were orthogonal, perhaps reflective of an orthogonal nature to the drivers of these patterns.

Four circadian phase modules of genes were identified by projection which then served as input to a promoter sequence analysis approach in *Arabidopsis*. We used a decision-tree based approach to search for combinations of elements which best classify genes into these modules. We identified a number of potential known and novel circadian *cis* elements which appear to act in conjunction with each other, and a classifier, trained using these promoter sequence motifs, resulted in high precision and recall. The groupings of these sequences changes in a consistent, progressive manner from one phase module group to the next suggesting that they may act in a simple additive manner to determine the final phase. We, therefore, infer what we propose is a framework defining the global structure of the plant circadian network.

introduce the wider concept of prediction of gene expression patterns based on sequence. also mention why an unbiased method would be preferred for our search for cis elements.

3.2 RESULTS

3.2.1 PATTERN INFERENCE THROUGH ICA

In order to better infer the network topology of circadian output genes, a combinatorial analysis of potential promoter elements determining phase-of-expression was carried out. The foundation step was the development of an unbiased method to identify circadian patterns of gene expression from within relatively short time-course transcriptomic data. We argued that this gives rise to a strong alternative the current favoured method of looking for preconceived patterns of gene expression where learning is strict on data. This method aimed to identify circadian patterns from observations made on the data itself. To do so, we employed independent component analysis (ICA) to identify dominant descriptive components or trends which we have termed "eigentrends", within data followed by projection of all data onto these dominant eigentrends, as portrayed on Figure 3.1.The hypothesis was that, in the absence of variation in external stimuli, variation in gene expression over a multi-day, free running time-course would be entirely due to endogenous processes within the organism. It became evident that the dominant endogenous processes would be circadian and these circadian patterns should, therefore, constitute the eigentrends which account for the most variation within the data.

A component analysis approach begins by identifying the first component, the vector which accounts for as much of the variability in the data as possible. It then looks for additional component. Each succeeding component in turn accounts for the highest variance possible within the data. A classical component analysis technique for detecting and visualizing relevant information from measured data is principal component (PCA)

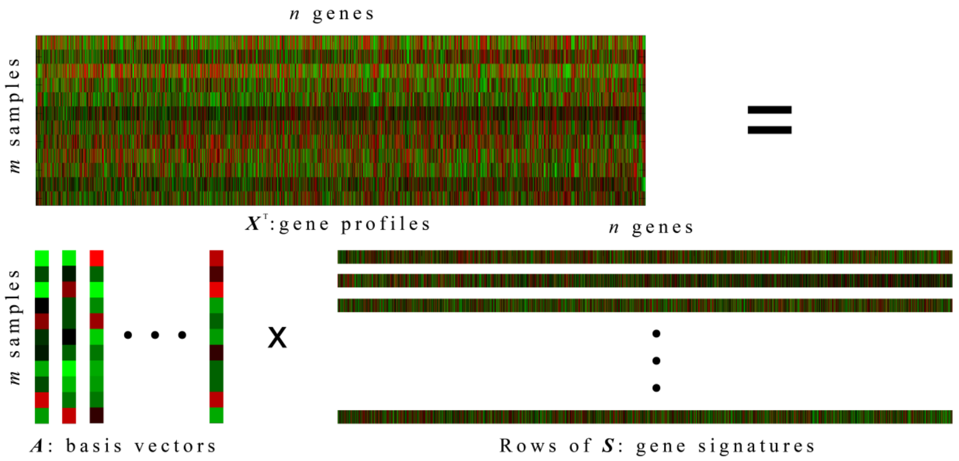


FIGURE 3.1: ICA decomposes the circadian microarray gene expression data. Here columns of a A and rows of S are separated for a clear visibility (green denotes low expresion whereas red denotes high expression).

(Alter, 2000). However, the fact that PCA necessarily identifies orthogonal components (uncorrelated with the preceding components) is problematic. In looking for circadian patterns of expression, we wished to remove this restriction as only two cycling orthogonal components with the same period are mathematically possible (essentially, sine and cosine patterns). Instead, we used ICA, a variation on PCA which does not impose the limit of orthogonality (Holter et al., 2001).

We began by analysing a well-characterised microarray dataset from the model plant, *Arabidopsis* (Covington et al., 2008). Plants had been entrained in light/dark cycles prior to release into constant light. The dataset comprised roughly 22,810 transcripts monitored for 12 time points taken at 4 hour intervals. For practical reasons we used the degree of kurtosis to sort the eigentrends (Liebermeister, 2002). We reasoned that latent trends with the most negative kurtosis can give us the most relevant information on the basis that more data points that expected will be positively or negatively correlated with these eigentrends. After ordering the eigentrends by kurtosis, the two most significant eigentrends were apparently circadian supporting this proposal, while the third possibly represents a damping rhythm (Figure 3.2A and B). A damping rhythm is indicative of a rhythm established under entraining conditions that damps following transfer to constant conditions as seen on Figure 3.4. Another standing explanation could be a general decline in expression levels with hints of a regular pattern. Subsequent eigentrends beyond the third showed an apparently random pattern.

Following an ordering of eigentrends by kurtosis, we determined how many were truly significant in terms of representing the data. The data was projected onto a subspace defined by the eigentrends and then compared to the original data by measuring the distance between the original data and its estimate based on these components.

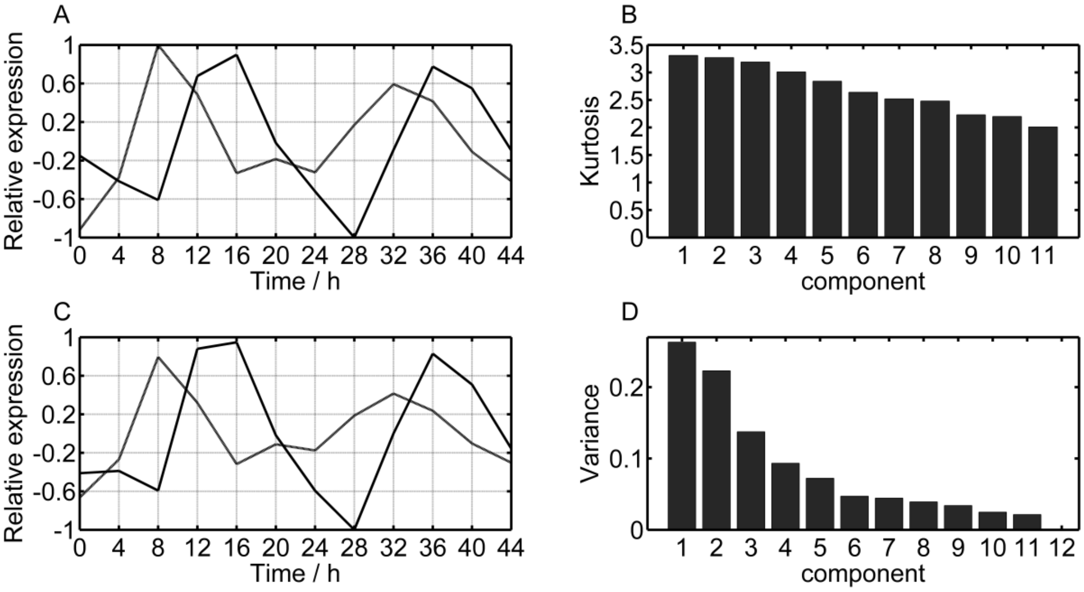


FIGURE 3.2: ICA and PCA performed on *Arabidopsis* circadian microarray time series. (A) First (peak at 8 hours) and second (peak at 12-16 hours) independent components plotted against time since subjective dawn. (B) The kurtosis of a frequency distribution of data projected along each independent component. (C). First (peak at 8 hours) and second (peak at 12-16 hours) principal components. (D). Variance within the data captured by each principal component.

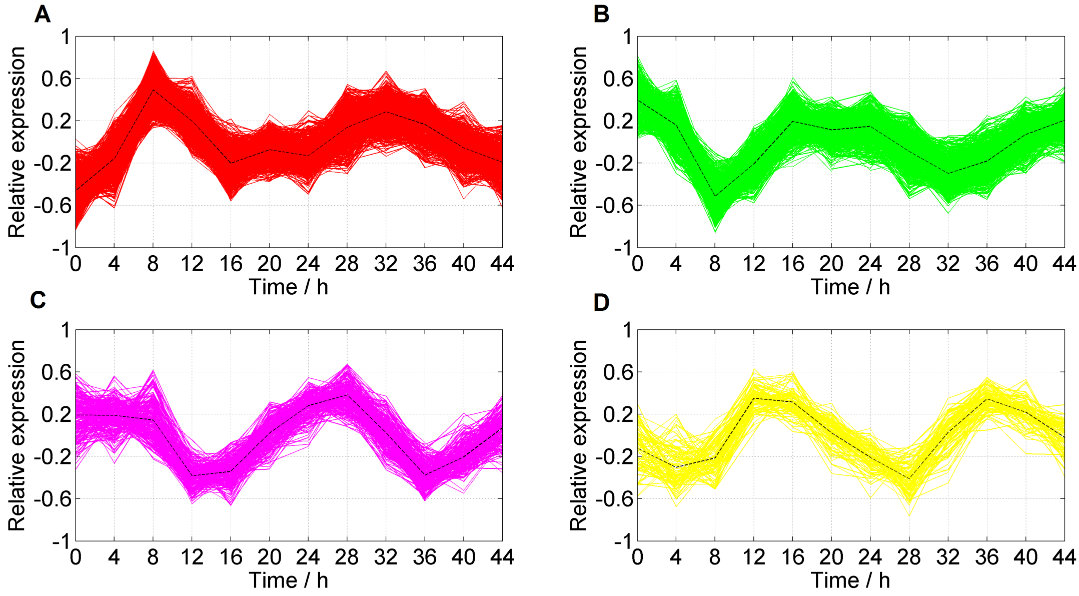


FIGURE 3.3: Average expression per detected module. The average expression per the four detected eigengroupings is depicted in the 4 subgraphs.

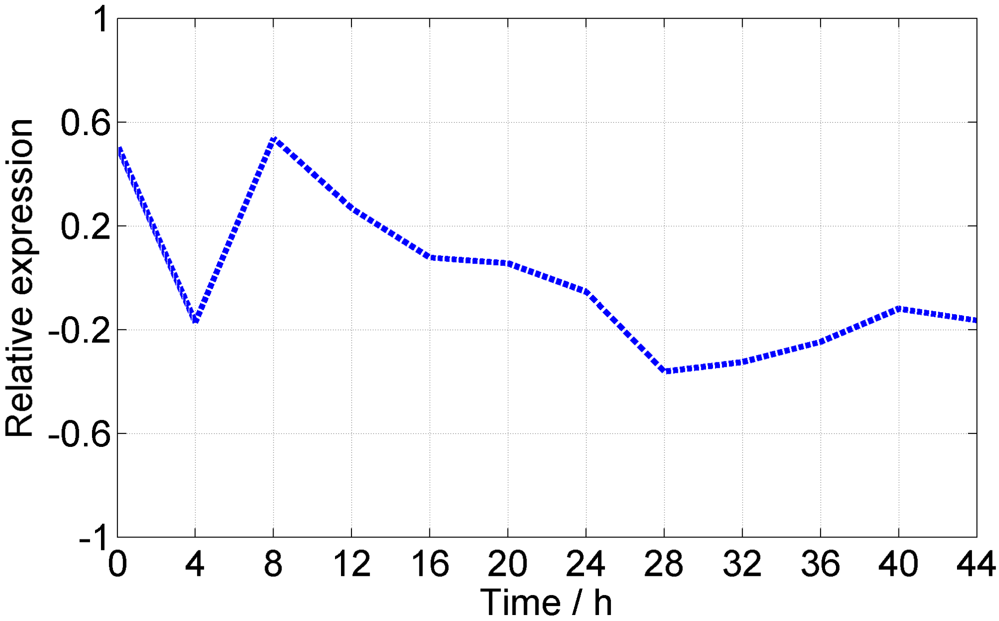


Figure 3.4: Dampening component. The figure portrays the damping rhythm is indicative of a rhythm established under entraining conditions that damps following transfer to constant conditions

This was done initially for the first eigentrend then recalculated as each subsequent eigentrends was added. We observed that the subsequent addition of eigentrends after the first three did not strongly decrease reconstruction error, suggesting that we have three significant eigentrends (Figure 3.5). We also devised an empirical significance test to determine the number of eigentrends which are significant in terms of representing the whole microarray data. Here ICA was carried out and reconstruction errors were calculated for 1000 datasets where the expression values within each gene's expression pattern were randomised. For each possible number of components we calculated the frequency that the randomised reconstruction errors were smaller than the calculated reconstruction error for the actual data. This approach also demonstrated that the first three eigentrends significantly described the data, while one, two, four or five eigentrends did not (Figure 3.5). Thus, ICA led to delineation of three apparently biologically relevant trends. The first two of these are "circadiantrends"; the third component, possibly representing a dampening rhythm, typical of rhythms which lose rhythmicity following transfer from driven to free-running conditions. The two "circadiantrends" delineated by ICA correspond to 4 phases seen in Figure 3.4 (positive and negative impressions of the curves) having approximately 4 hour / 8 hour difference.

On comparing the results with a PCA analysis of the same data we observed that the two apparently-circadian trends identified by ICA are also remarkable in almost exactly concurring with the first two eigentrends identified by PCA (Figure 3.2C and D). The concurrence was confirmed by correlation analysis, giving an R value of 0.9987 and 0.9972

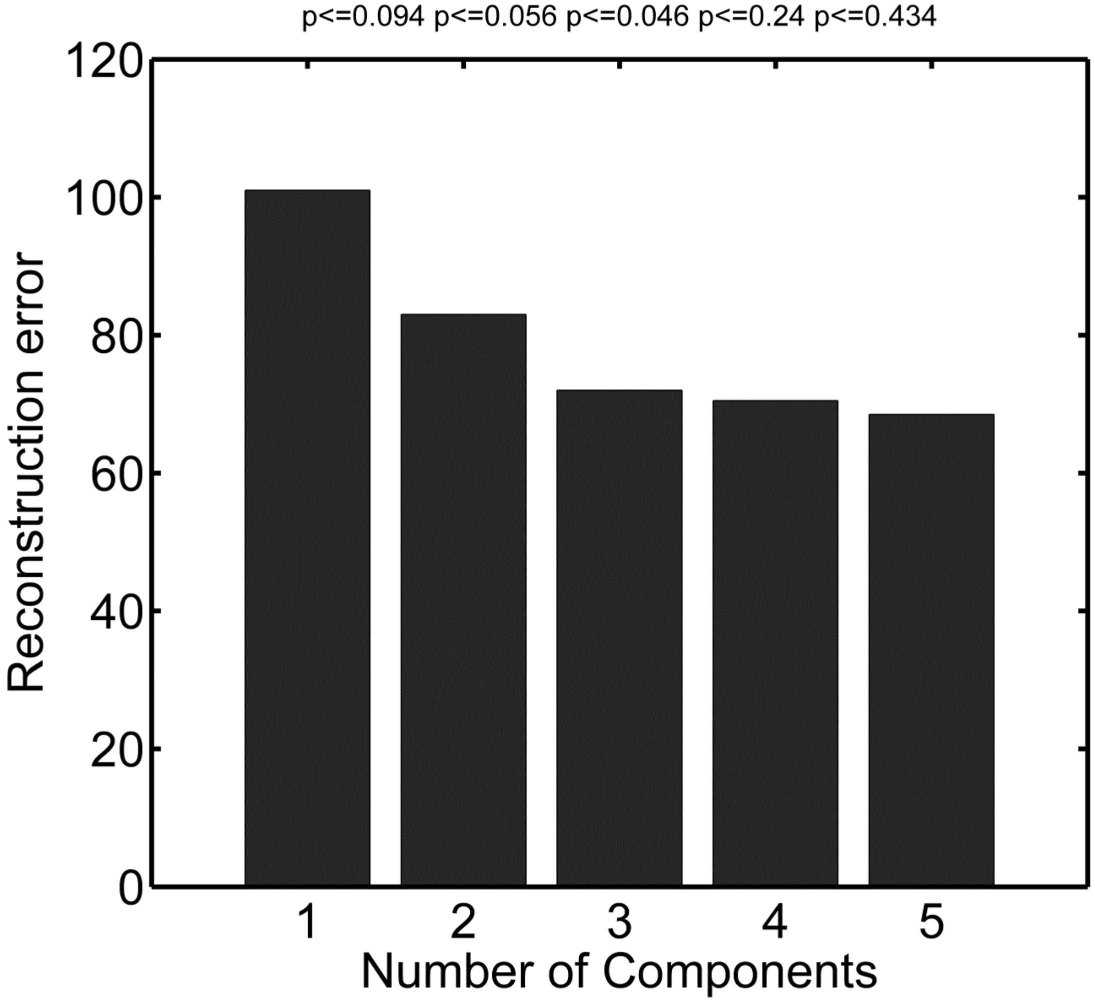


FIGURE 3.5: Reconstruction error for ICA components of an *Arabidopsis* circadian microarray time series. After ordering by kurtosis, variance reconstruction error was plotted for addition of successive independent components. P values are shown for an empirical significance test to determine the number of independent components which are significant in terms of representing the whole microarray data.

for the first and second eigentrends respectively. This agreement between results of ICA and PCA indicates that the two circadian eigentrends identified by ICA are orthogonal despite this not being imposed by the method. We propose that this mathematical orthogonality could possibly represent an underlying biological orthogonality in the way that rhythms are generated at the molecular level.

3.2.2 ASSIGNMENT OF CIRCADIAN GENES THROUGH PROJECTION INTO 2-DIMENSIONAL SUBSPACE

Having identified two dominant circadian eigentrends and, therefore, four dominant oscillatory patterns, we sought to identify circadian genes themselves by sorting the data by similarity in expression to these eigentrends. This operated on the assumption that circadian genes will show strong dot product projection similarity to one or both of the two key circadian eigentrends that we identified and, thus, fall in a circular region some distance from the origin. This method was chosen in favour of a simple Pearson correlation. In a correlation scatter plot, the significance of genes with a low level of expression, whose pattern of expression might be considered as noise, can appear to be magnified. Effectively, the separation between signal and noise genes is decreased for a correlation vs. a projection scatter plot.Moreover that gave rise to groupings as seen in Figure 3.3. We imposed a cut-off at a distance of 0.8. Using this approach we classified 2959 genes as circadian (Figure 3.6A and S3). Effectively, the radial coordinate constitutes a measure of cyclicity. This is depicted in Figure 3.6A where the outermost genes are those we define as circadian genes. Effectively, the angular position around the plot represents the phase of expression.

As expected this method identified circadian genes defining the full range of possible phases. We noticed a distinct clustering for genes showing high dot product with the first component on the x-axis (Figure 3.6A). This would suggest that the most prevalent phase of peak expression is just after dawn. This subset was inspected for known core circadian genes and for all circadian genes identified as associated with the circadian clock according to the *Arabidopsis* Information Resource Gene Ontology database. All core genes examined were found to be present in the circadian subset. Core genes include CIRCADIAN CLOCK-ASSOCIATED (*CCA1*), LATE ELONGATED HYPOCOTYL (*LHY*), PSEUDORESPONSE REGULATOR (*PRR*) 3, 7 and 9, TIMING OF CAB EXPRESSION 1 (*TOC1*), LUX ARRHYTHMO (*LUX*), and EARLY FLOWERING 3 (*ELF3*). Furthermore, *CCA1* and *LHY* were found to be in antiphase with *TOC1* as indicated by the opposing positions in the agreement with their known antiphasic pattern of expression (Figure 3.6A). We also noticed that the core clock genes tended to be found directly close to the axis representing the second component whereas the remaining population of circadian regulated genes were found spread around the circadian subspace with a clustering around the axis of the first component, approximately six hours behind these effectors. This perhaps adds weight to the speculation that there is some degree of biological orthogonality represented by the two key orthogonal eigentrends which define these axes.

Our identification of 2959 circadian genes within the Covington dataset is comparable to the 2897 identified by Covington et al. (2008) using the regression-based COSOPT method (Straume, 2004). However, our method identified only 1168 genes in common with the COSOPT method highlighting the fact that this approach is quite distinct from the most commonly used biased approach to circadian microarray analysis.The identification of circadian genes is very much dependent on the method used and, given the range of methods available, no benchmark set of Arabidopsis circadian genes has been

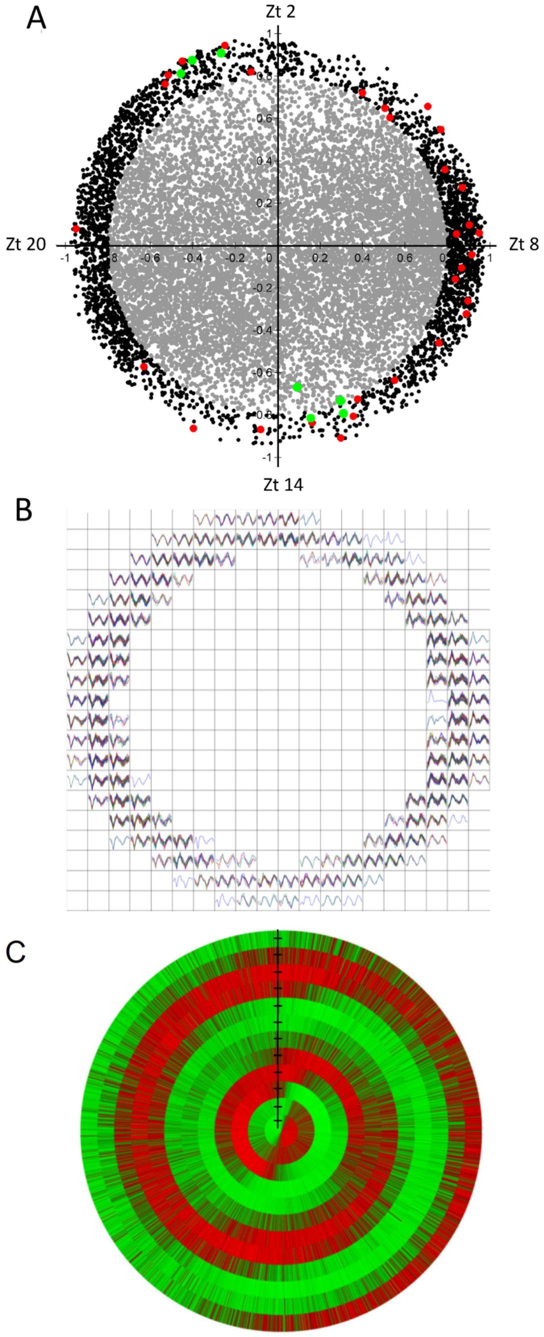


Figure 3.6: (A) Strong circadian genes identified by projection onto ICA components. Expression patterns of all genes from *Arabidopsis* circadian timecourse data were projected onto the first (x-axis) and second (y-axis) ICA components. Black dots represent circadian genes, defined as lying at a distance of at least 0.8 from the origin. Grey dots represent those classed as non-circadian. Green dots represent the genes of the central clock loop (*CCA1, LHY, TOC1, PRR9, ELF3, ELF4 and LUX*). Red dots represent other oscillator-associated genes. Zt: Zeitgeber time. (B) Expression profiles of circadian genes. The projection scatter plot was divided into a 20 by 20 grid and all the genes in each grid box are displayed together as line graphs. (C) Circular heat map of gene expression in which angular positions of genes were used to order rows. Each row from the outside in represents the data-points for a single gene across the circadian experiment. The circular map shows normalized gene expression patterns. Red: high expression; green low expression.

defined for comparison. To demonstrate the low agreement between the methodologies Venn diagram is presented in Figure 3.6 which portrays the overlap between the genes detected by Lomb Scargle, Haystack and ICA. Strikingly low is the overlap between all 3 methods, yet this is a finding consistent with previous studies. As the plant circadian benchmark set per se does not exist biologically speaking, it is incredibly difficult to establish the ground truth set. Remaining comparisons were deposited in the Appendix. The genuine oscillatory pattern of our selected genes is, however, demonstrated in Figure 3.7 (B and C). Two approaches were used to demonstrate this. Firstly, the changing oscillatory patterns of the individual genes plotted around the projection are presented in Figure 3.6B. Secondly, genes were ordered through angular positions in the ICA result around a circular heat map (Figure 3.7C). These methodologies also portray the inherent topology of the data as a continuous distribution. This further confirms the advantage of this method as a non-biased approach for the identification of circadian genes. Here, clustering, an alternative non-biased approach would force cluster topology on a continuous distribution preventing characterisation of individual genes along such a continuous distribution. Our method, furthermore, adds the ability to order circadian genes by more than just peak phase as has been used in the case of biased pattern-matching approaches. Genes sharing peaks but potentially having quite different patterns of expression in terms of peak shape, could not be ordered solely on the basis of peak time, whereas our method also allows such additional information to be considered. This is particularly significant given the discrete nature of the data with samples taken only every four hours. These approaches portray the inherent topology of the data as a continuous distribution. This further confirms the suitability of this method as a non-biased approach for the identification of circadian genes. Alternative non-biased approaches using a classification method that groups genes according to their co-location in the neighbourhood of a point, like k-means clustering, would not be appropriate for dealing with ring-like distributions. Similarly, hierarchical clustering through average linkage nicely portrays the close relations of genes within clusters yet the complete ordering is lost. In other words here the cluster like approach would force cluster topology on continuous distribution.

3.2.3 ROBUSTNESS OF THE CIRCADIAN CLOCK GENE EXPRESSION MODULES

We also applied this method for identification of circadian genes to additional datasets. We chose a second dataset for Arabidopsis and additional datasets for Zea *mays*, and

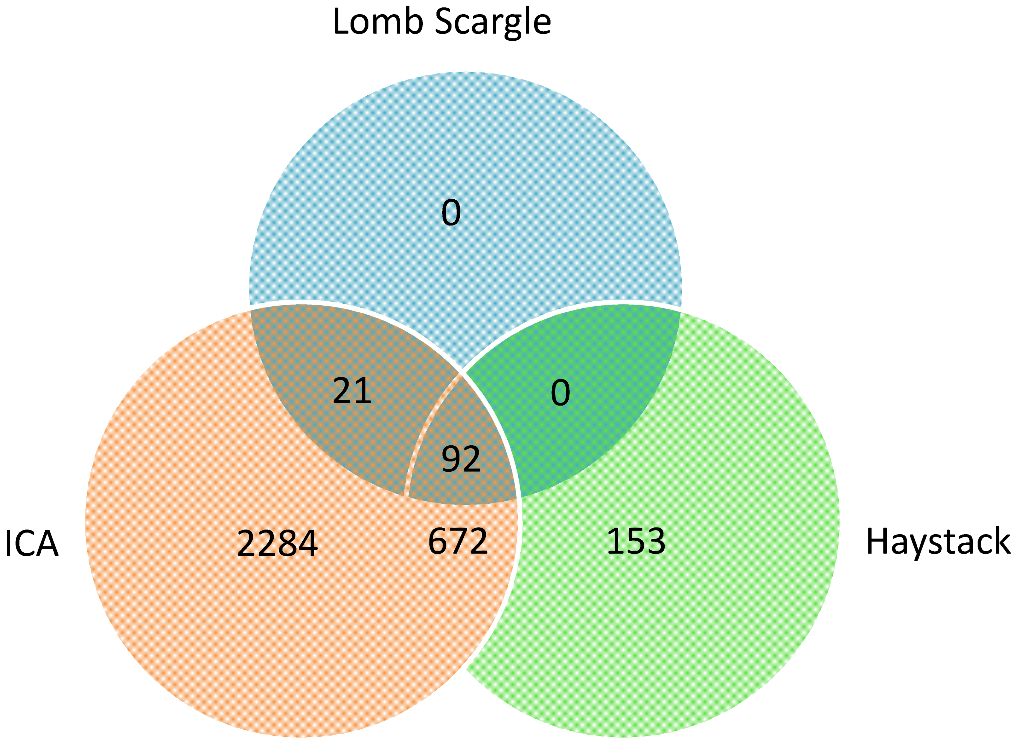


Figure 3.7: Venn diagram portrays the overlap between three methods for selecting circadian genes including Lomb-Scargle, Haystack and ICA.

Oryza *sativa*, all generated using the same entraining and free running conditions (Edwards et al., 2010, Hayes et al., 2010, Xu et al., 2011). By repeating the discussed workflow we found that the pattern of eigentrends was highly preserved in another dataset for Arabidopsis and in other species as shown in Figure 3.8. It shows the profound conservation of orthologues through inspection of the angular positions of genes across species. In all cases, there were two dominant eigentrends which were cyclic in nature with a periodicity close to 24 hours. These same circadiantrends were identified using both ICA and PCA and this finding suggests that the two orthogonal eigentrends found in the circadian data in the first Arabidopsis sample are conserved throughout a range of plants, encompassing both the monocots and dicots, further suggestive of an underlying biological significance that is also conserved. Strikingly, the position of key circadian orthologues was also preserved in projection scatter plots (data not shown). This is consistent with previous analysis which has shown that circadian orthologues correlate very well between different plant species and are, thus, expressed at approximately the same phase (Priest et al., 2009).

We also confirmed conservation of the wider patterns themselves across species using comparative mathematical framework. Orthologues were found between *O. sativa* and Arabidopsis. Altogether 2209 orthologues were found for *O. sativa*. Cross-species microarray comparisons are complicated by noise, assignment of homology, probe quality,

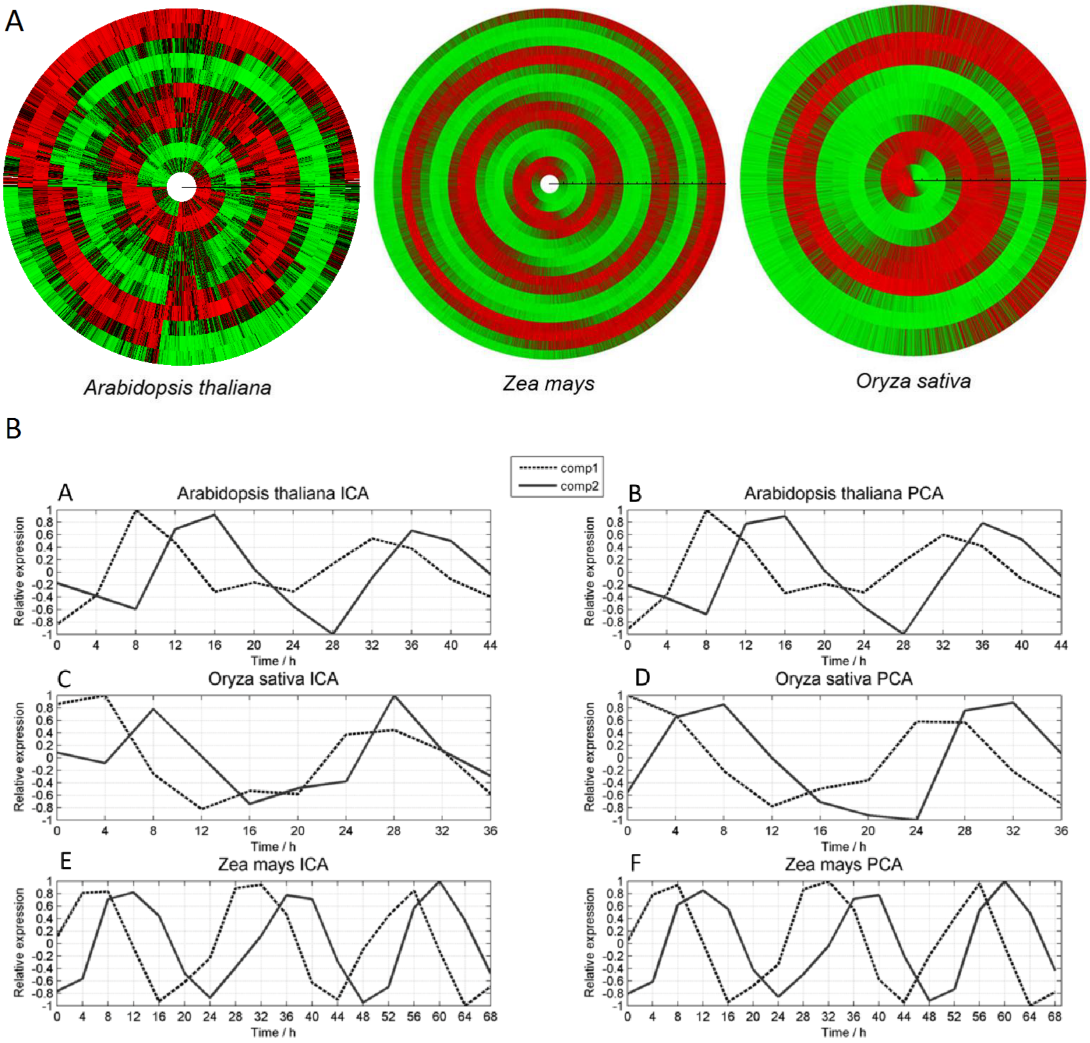


FIGURE 3.8: A. A circular map in which angular position of genes were used to order rows. The circular map shows normalized gene expression patterns of circadian transcripts across species, the color scale ranges from red (high) to green (low). The figure highlights the conserved phasing of circadian genes across species. B. It shows the dominant trends for circadian timecourse transcriptomic data across species presented via ICA (A, C and E) and PCA (B, D and F) plots. A second Arabidopsis transcriptomic data set (A and B) and datasets for *Oryza sativa* (C and D) and *Zea mays* (E and F) were analysed (leaf tissue). In all cases, the two, similar dominant independent components were the only circadian trends. Strong correlation is observed between ICA and PCA trends confirming that these two orthogonal circadian eigentrends underlie the circadian system across several plant species.

platform variations, laboratory effects, genetic background, dynamic environments and the status of the plant. However, we used correlation coefficient as a measure to infer that the transcriptional behaviour of circadian genes is highly conserved across species. This metric assumes that, whereas the overall raw expression values may vary between studies, the intergene correlations will be more invariant (Cope, 2011). We followed the method suggested by Doherty and Kay (2010). For all Arabidopsis circadian genes (GSE8365 dataset) with a circadian orthologue in rice, the correlation of each gene with its orthologue was determined, then a frequency distribution of these correlations was plotted (Figure 3.9). The distribution of correlation values for all circadian probes shows that orthologous genes have a much higher correlation compared to a negative control representing correlation between orthologous genes when data points for each gene was randomly shuffled (Figure 3.9). In the test data, 365 genes had corCor > 0.1; in the negative control no gene had corCor > 0.08. This clearly suggests than when we look at the data globally, taking the entire transcriptome into consideration, a large proportion of circadian orthologues are expressed in a conserved way.

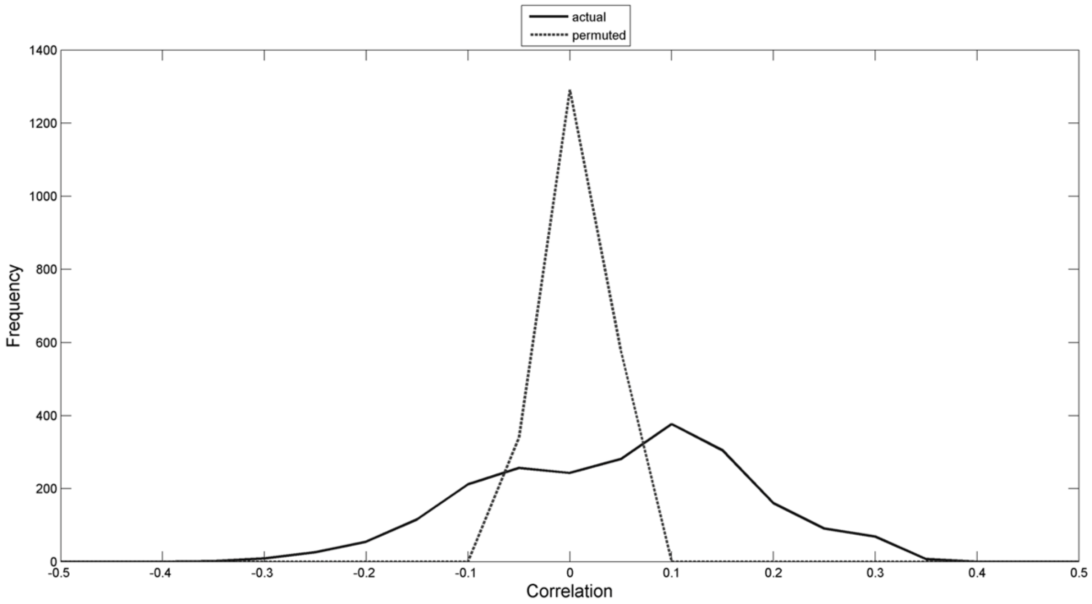


FIGURE 3.9: Integrative correlation coefficient between species. Distribution of correlation of correlation coefficient between Arabidopsis and Oryza sativa orthologous circadian genes. The orthologous circadian genes show significantly higher integrative correlations using actual observed data (black line) than the null distribution which was generated using permuted data where gene expression values within each gene were randomized (grey line).

The fact that individual gene patterns are conserved in addition to the conservation of the eigentrends suggests that the overall architecture of our circadian four "modules" is also preserved. That further opened novel avenues for research into the evolution of circadian networks. The significance of this modularity was assessed across species

ID	STD	Mean	Jaccard	Normalized x STD
class1	0.0037	0.0414	0.2716	61.0110
class2	0.0069	0.0208	0.2657	35.0070
class3	0.0017	0.0850	0.1758	53.5500
class4	0.0033	0.0633	0.1898	42.9600

Table 3.1: Comparison of phase module composition across species. The Jaccard similarity score, TP/(TP + FP + FN), was used to score the overlap between phase module compositions in Arabidopsis and Oryza *sativa* using orthologues identified as circadian in both. The, standard Deviation (STD) and mean of the distribution of Jaccard similarity scores on randomized module partitioning are displayed, as is the normalized Jaccard similarity score, that is, the number of standard deviations from the mean . Classes 1 and 2 correspond to subjective dawn and subjective dusk-phased genes respectively whereas classes 3 and 4 correspond to the middle of the subjective day and night respectively.

datasets using the Jaccard similarity score (Fortunato, 2010). The Jaccard coefficient measures similarity between sample sets (modules). It assesses the number of common members of two sets as a proportion of the total number of members in the two sets. This was used to score the overlap between two equivalent module compositions across Arabidopsis and Oryza *sativa*. Additionally, randomisations were used to define the significance of each specific score. This produced a normalised similarity score for each module, expressed as the number of standard deviations from the mean of the distribution of Jaccard similarity scores for equivalent randomised module structures (Table 3.1). For the four phase module classes the number of standard deviations from the mean was between 35 and 61 indicating an extremely significant conservation of the members of these four phase modules.

Evidence for the substructure (circadian architecture relying on the eigengenes) was determined and has been quantified using yet another method. Empirical p-values were derived for each of the eigenvalues corresponding to each of the eigenvectors across datasets. Significance analysis (that is to determine the likelihood of the eigengenes to be determined by chance) was performed for that purpose with calculation of null distribution. P values were derived empirically from the null hypotheses. A significant eigengene is one that represents a greater proportion of variation than expected by chance. The selected eigentrends have a highly significant p-value <0.001. That suggests that the points are overlaying each other on a circular plot and the results are not coincidental.

The statistical inference reliant upon random matrices provides another line of evidence for the eigentrends and outcome of analysis being the reflection of topology. Strong evidence should translate into direct interpretability of the selected components.

Furthermore to eliminate doubts, as in the components serving as building blocks, the eigenvalues of circadian eigentrends within Arabidopsis were assessed using a comparative mathematical framework. The antisymmetric angular distance between the datasets (two different Arabidopsis datasets used that is Covington and Edwards) was used to assess the equal significance in Arabidopsis, that is the respective eigentrends have identical phase angle in both Arabidopsis datasets. An angular distance of 0 would be indicative of significance whereas a lack of significance would be determined through a distance of $\pm \pi/4$, that is for both the eigenvalues and eigengenes. The findings further supports the notion of conserved architecture across species as the first 2 components have an angular distance close to 0. That concludes the conservation of topology aspect of the analysis.

The different types of biological interactions provide distinct, yet complementary, insights into cellular structure and function. A key opportunity lies in reconciling these complementary network views of the cell into cohesive models. Protein-protein, transcriptional, genetic, metabolic or other types of molecular networks can be combined for exploration and validation. One option involves the identification of overlapping clusters of physical and genetic interactions; these joint modules implicate genes acting within a pathway. Here the benefit was taken from the Arabidopsis Interactome Mapping Consortium which has mapped roughly 6,200 reliable interactions between about 2,700 proteins using yeast two-hybrid-based system. The objective was to compare distribution of correlation between all possible interactions of 196 genes (limited to TF's that were found circadian) and then to select an equal number of random correlations in Arabidopsis *thaliana*. Essentially the network obtained from these gene pairs was overlaid with the true calls derived from the protein interactions confirmed in Arabidopsis. This overlay serves another source of evidence for the true existence of the links and the distribution of values across two datasets can be seen in Figure 3.10. In the present work the hypothesis was that more such PPI interactions will be detected in the real versus random network. That is was the results have shown too. A similar methodology was attempted in this work using the BIOGRID interaction. The advantage of this specific method is its comprehensiveness (one study) whereas interactions in compiled databases come from several different experiments. The agreement between both networks suggests the greater likelihood of true associations whereas the lack of agreement could mean either of two possibilities: a false positive or a different mode mechanism of action.

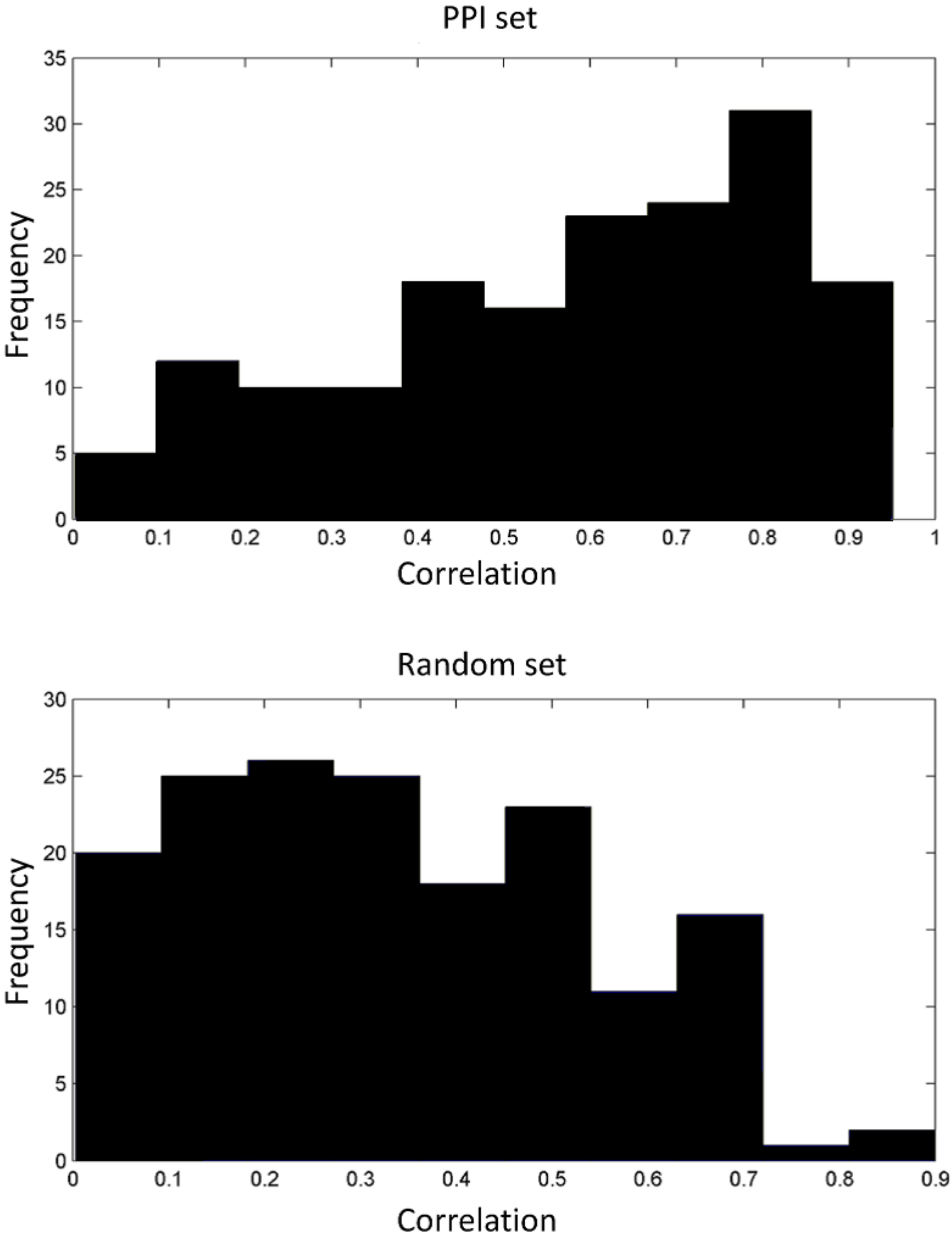


FIGURE 3.10: The compared distribution of correlation between all possible interactions of 196 genes and select equal number of random correlations in Arabidopsis *thaliana* (top. PPI vs coexpression) (bottom PPI vs random correlations derived from a shuffle matrix)

3.3 DISCUSSION

We set out to better define the network architecture of the plant circadian clock. Our primary aim was to delineate the modules with the ultimate objective being a combinatorial approach to identify potential promoter elements defining the phase of expression of circadian genes. Previous research looking for phase-specific promoter elements has focused on the identification of overrepresented single elements (Harmer, 2000, Michael et al., 2008), but the expression patterns of a large proportion of clock output genes remains unexplained.

Modules are the functional building blocks of the system. Uncovering the circadian genes and the modules they aggregate into is the initial step. We aimed to test the hypothesis that progressive combinations multiple promoter elements acting in concert of may be responsible for the full range of phases observed in plant circadian output genes. To do so the initial aim was defining the set of circadian genes. A dynamic system like the plant circadian clock which arises from gene regulatory networks has highly dimensional state spaces and is highly dependent upon large number of parameters. To make sense of this highly complex system one has to infer points of regulation and how the persistence of the clock (robustness versus fragility) is maintained to allow for adaptability. Hence the persistence, strength and timing of particular events has to be uncovered. For the purpose the boundary between the circadian and the non-circadian genes has to be established.

Analysis of the global gene expression data has commonly been used with the aim of understanding the properties of the plant circadian system. These approaches have allowed individual circadian promoter elements to be identified but the global topology of the system as a whole remains elusive. To attempt to better describe the circadian network, a combinatorial approach was used for the identification of promoter elements. A novel, non-biased method for the identification of circadian genes was used, allowing reliable selection of intrinsic cyclic patterns of expression.

This non-biased approach identified two dominant, inherent circadian trends underlying the data. Surprisingly, these proved to be orthogonal, exhibiting a 90 degree angle between them. Furthermore, these trends were highly conserved across several plant species. These orthogonal trends represent an orthogonal nature to the drivers of the network. Four phase-specific modules of circadian genes were generated by projection onto these trends. In order to identify potential combinatorial promoter elements that might classify genes into these groups, a random forest pipeline was employed which merged data from multiple decision trees looking for presence or absence of element combinations. That led to identification of a number of regulatory motifs which aggregate into coherent clusters capable of predicting the inclusion of genes within each phase module with very high fidelity. These motif combinations change in a consistent, progressive manner from one phase module group to the next, providing for the first time a potential global description of the topology of the plant circadian system. In essence stringent methods will commit false negatives whereas liberal methods will generate false positives. The power of those tests is difficult to assess without having a benchmark hence we propose resorting to learning on the data. Several method were presented all approaching the problem of signal detection from a slightly a different angle. Two in depth reviews particularly focus on their advantages, Dequeant et al., 2008 and Zhao et al., 2008. These yet use a subset of known for a given problem genes as the benchmark, unknowingly of the size of the target group. Recently, a study by Deckard et al., 2013 evaluated several methods to detect periodic signals yet with the advantage of using a synthetic dataset as the ground truth. The synthetic dataset consisted of both non periodic (flat and linear) and periodic profiles (cosine, 2 cosines with different periods and amplitudes, cosine damped, cosine peaked, cosine with linear trend and cosine with exponential trend). Additionally different numbers of samples were tested and the Gaussian noise varied. The ability of these methods to recover period, phase shifts and amplitude was thoroughly tested. The findings suggest that curve shape has the largest impact on the scoring of biological signals by the tested periodicity detection algorithms. Algorithms such as Lomb Scargle, De Lichtenberg, Persistent Homology and JTK cycle rely on comparing data with reference curves such as sinusoidal curve and a user-specified curve therefore, they will perform most accurately when the data match the assumptions specified by the reference curves. The scores and the data are often difficult to compare due to nature of the data and algorithm. The discussed methods have different yet at times not overlapping advantages. They define periodicity differently, weigh components of rhythmicity differently, respond to noise differently, at times allow for irregular intervals and missing data. It is clear that at times amplitude could be a useful measure of regulation, whereas at others phase shift which is important measure of timing could be of importance. Being able to accurately estimate phase shifts between transcripts can allow one to reconstruct timing and suggest regulatory relationships. All these methods in a way call for a supervised search rather than pure unsupervised learning.

3.3.1 ICA AS A METHOD OF IDENTIFYING CIRCADIAN GENES IN SHORT TIMECOURSE MICROARRAY DATA

Identification of meaningful phases-specific groups of circadian genes is an essential prerequisite to analysis of promoter elements. Previously-used algorithms have included robust periodicity testing, Bayesian mixture models, matching to model functions, Fisher's G testing, the Lomb Scargle periodogram, Fourier transformation, and the Laplace periodogram among others employed (Doherty and Kay, 2010). Conceptually, the methods encompass two primary categories: pattern matching in the time domain; or signal decomposition or filtering in the frequency domain. In the case of pattern matching approaches, the patterns are predetermined and, therefore, biased. As such they may not truly reflect the dominant expression patterns among the oscillating genes. Signal decomposition methods look for frequencies of approximately 24 hours in an unbiased manner but their accuracy is strongly linked to the duration of the timecourse data available. For short timecourses of two days which are commonly used for circadian microarray experiments, these methods have limited power to identify circadian genes with any certainty. These methods to identify circadian genes also suffer from the additional drawback that they do not naturally sort genes into phase groups; instead, phase groups for subsequent analysis are usually imposed artificially.

We identified a novel, unbiased method for the identification of genes showing a circadian pattern of expression from short time-course, relatively low-resolution microarray data using a method of ICA to identify components in global gene expression. We followed this by projection of individual gene expression patterns onto these components. Significantly, ICA identified two orthogonal components which accounted for the majority of variance in the global gene expression data. Projection of individual gene expression data onto these components showed that known circadian clock-regulated genes showed a circular distribution around the edge of the scatter plot, correlating well to one or both of the components. This method forms a powerful approach for the identification of circadian genes but also naturally sorts circadian genes into phase groups according to closeness to one of these inherent components.

It is tempting to speculate that the two orthogonal components identified may be indicative of some underlying biological orthogonality inherent in the way that plant rhythms are generated. One can conjecture that each eigentrend represents a regulatory phase pattern that is biologically interpretable. In a nutshell it is accepted that the clock operates through the cooperative relationship between ”morning” genes and ”evening” genes. Following their morning peak, expression of the morning genes is suppressed during the day by a group of pseudo-response regulator proteins acting sequentially.

Expression of these pseudo-response regulators is, in turn, suppressed by evening complex proteins which, therefore, releases the repression of the morning genes. As morning gene proteins re-accumulate they repress the evening genes allowing the pseudo response regulators to accumulate and to begin repressing expression of the morning genes once again. These morning and evening groups of central clock genes also regulate output genes suggesting that, essentially, the plant clock consists of two main gene expression modules. By contrast, our findings suggest four distinct modules, groups of genes showing either positive or negative correlation to the two dominant eigentrends. One of our dominant trends corresponds well to the phase of expression of the morning and evening genes of the central clock along its positive and negative axes. It is possible that output targets in these modules may be regulated in the same way as the morning and evening genes of the central clock. This would still leave two other significant modules unexplained. One enticing possibility is that this may represent a significant number of output genes regulated by different combinations of the same promoter elements having the net effect of an intermediate peak time.

The extent of conservation of our four modules across species is also striking both in terms of the patterns and of the constituents if the modules. This adds further significance to the observation. Orthologues of the vast majority of clock genes in Arabidopsis have been isolated in a wide range of species including rice and maize suggesting a common mechanism driving rhythmicity. Capturing the module-module interactions also, therefore, takes on a higher priority as it will likely allow any findings to be applied across the plant kingdom.

3.3.2 CONSERVATION OF COMPONENTS

The extent of conservation of our four modules across species is striking. Orthologues of the vast majority of clock genes in Arabidopsis have been isolated in a wide range of species including rice and maize. Their cycling resembles that of Arabidopsis. Conservation of expression patterns of individual circadian orthologues has been shown previously cite Michael. However, here, these findings are confirmed on a large scale using an integrative correlation coefficient. We showed that for orthologues identified as circadian using our approach, their patterns of expression were much more highly correlated than would be expected by chance. Yet what is significantly more interesting is the conservation of the modules. We used the Jaccard coefficient to measure similarity between module sample sets across species. These findings demonstrated extremely significant overlap between the module samples sets in Arabidopsis and Oryza. As such, these modules may be very significant on a wider level throughout plant biology. Capturing the module-module interactions also, therefore, takes on a higher priority as it will likely allow one to infer the dynamic changes that maintain the system in operation. PPI overlays were previously done to identified protein complexes through the search of regulatory and PPI data jointly in yeast (Tan et al., 2007).Protein-DNA interactions have also been combined with metabolic networks to delineate the effects of transcriptional regulation on biochemical processes.

3.4 CONCLUSION

The findings described here demonstrate the applicability of a novel ICA based approach for the identification of circadian genes in short time-course microarray data. The method revealed hitherto undisclosed characteristics of the plant ”circadiome”, suggesting the existence of two orthogonal biological trends underlying the patterns of expression data produced and this approach has also revealed a high level of conservation of phase modules within the circadian gene expression patterns of distinct plant species. However, such an approach could equally be applied to data generated from any biological sample. In our case this has provided a source of new putative mechanisms which may underlie the organisation of the plant circadian system. This system would also be equally applicable throughout biology to any classification system used to group genes on the basis of expression pattern. Results can be extrapolated to crops of great agronomic importance. Essentially, we find a basis that is a linear combination of original data, best re-expresses and project circadian data. This basis is more informative in terms of the understanding of architecture and allows for the detection of driver nodes and delineation of the communications between them than fitting to presumed models.

Chapter 4

THE RANDOM FOREST AND BEYOND

4.1 INTRODUCTION

A major aim of circadian biology is to be able to predict the expression of genes given their regulatory sequences similarly as it is possible to predict the protein sequence from the open reading frame. To what extent can we predict the phase of expression of genes given their sequence? Can we predict classes of genes given the attributes of sequence motifs and information pertaining to their location and frequency? Are these attributes independent and if so, how so? These questions are fundamental to the understanding of transcriptional regulation and have been put forward and explored from a circadian perspective.

Machine learning, a branch of artificial intelligence, is about the construction and study of systems, about learning from data. One particular type of learning is reliant upon decision trees. This Chapter will introduce how the ensemble learning random forest methodology was applied to answer the fundamental question that is raised, that is whether we can delineate clusters motifs that will be able to predict specific phases with high accuracy for particular phases of interest. Given features (motifs at certain positions) can we predict classes (gene expression patterns) of genes? That can be translated into classical machine learning problem: can we predict classes given the features. Here the features are primarily *cis* motifs, but several other designs including network properties were analysed. Such *cis*-regulatory sequences direct cell-specific and inducible gene expression in response to signal transduction networks, and their value has been shown particularly when SNPs occurring at these sites lead to aberrations. Here the features, the attributes and the classes are defined as follows:

The *cis*-regulatory sequence is a noncoding sequence that controls the expression of genes within specific cells, at specific amounts, and in response to specific stimuli. These include promoters, enhancers, silencers, and insulators. These further affect the magnitude, the timing, and the cell specificity of gene expression. The abundance of both computationally and experimentally derived sequence elements and their growing usefulness in defining genetic regulatory networks and deciphering the regulatory program of specific genes makes them important tool for the understanding of the plant clock and its outputs. Simply considering the length and the degeneracy, the probability that a random 6-mer matches a specific hexamer binding site is $(1/4)^6$, so the site occurs about once every 4^6 (= 4,096) bp in a random DNA sequence. The potential justifications for the deviation from the expectations upon particular motifs will be discussed later in this chapter. Sequences are often represented by the sequence where each consensus motif is scaled with the information content of the base frequencies at that position. Position weight matrices can then be used to calculate the specific-binding free energy (relative to random background DNA) of a given sequence.

Transcription factors are proteins that bind *cis*-regulatory sequences. Once bound, TFs can influence RNApolII activity as well as modulating chromatin remodeling events. Often multiple TFs can activate a target promoter, with both OR-like logic, that is 1 TF in play or AND-like logic (calling for several TFs) regulatory architectures. Activators are TFs that activate transcription whereas repressors inhibit transcription.

A promoter is a sequence of DNA next to the transcriptional start site of a specific gene which serves as the initial binding site. It can be divided into the core, proximal and distal sections. Promoters do control the expression of a gene in response to the TFs. Several proximal promoter binding TFs are organism specific, for example the GC box in mammals, the MIG1 site in yeast, and the Y-patch in plants. Understanding the relation between the function of the promoter and its architecture will for example shed light into how the circadian clock operates. Sites can occur upstream being the negative base pair coordinates and downstream being the positive base pair coordinates from the transcription start site.

Plants have many transcription factors, and Arabidopsis has a compact genome giving rise to promoters with dense clusters of *cis*-regulatory motifs 500 bp upstream of the transcription start site. Many classical methods to reconstruct such regulation focused on Boolean logic, yet these do not cover the range of phenotypes that cannot be modeled by simple binary functions. Parameterization of logic can indeed affect the intuitive design of promoters. In the present analysis it serves to design the best matrices and the performance of the classifier of central interest. Motifs are typically represented by strings, strings with mismatches and position weight matrices. A number of motif discovery tools were developed over the last decade including Gibbs Sampler, MEME, Weeder, Improbizer, DME, DEME, A-GLAM, RED, AlignACE, ELEMENT (Bailey et al., 2006, Bussemaker et al., 2001, Hughes et al., 2000, Mockler et al., 2007, Tompa et al., 2005). There are several key methods for recognizing those regulatory motifs within promoter sequences like forming a position weight matrix from experimentally confirmed binding sites; word frequency analysis of short sequences at each promoter position; and correlation of motif presence with similar expression profiles to name some core strategies. AlignACE for example finds multiple motifs in any given set of DNA input sequences. The motifs are defined relying on base frequency patterns of the information rich columns of the aligned sites. It relies upon Gibbs sampling algorithm. The list of potential tools is elaborate and many of these are reliant upon each other. They differ by the learning principle employed to infer the model parameters and by their capability of learning the position distribution of the binding sites. For computational approaches, the fundamental improvements include searching for differentially abundant motifs and learning a position distribution. DISPOM uses discriminative learning principle and the position can be learned from the data and hence was used in this study. The information about motifs in promoters can be represented in the form of a matrix where each element annotates the number of occurrences of motif in the promoter of target genes. Many overrepresented motifs in the promoters were identified using several algorithms operating on different principles. The upstream regions were used for computation of significance compared to random genes. The larger the significance value, the higher its statistical significance. Matrices were constructed relying on overrepresentation analysis and relying on all combinations of 8mers of letters A,T,C,G. Altogether, 48 features were preselected (Appendix). Coverage indicated the percentage of genes that have at least one occurrence of a particular *cis* motif. A series of statistical tests were carried out to determine the significance of a given motif, including classic and novel hypothesis testing procedures. Decision trees which became the method of choice in this study have been widely adopted in biology from protein function prediction to molecular splicing. That is not to say that other methods were not tested, to the contrary they were and the results were compared; nevertheless, that will not be the focal point. Many types of conundrums would benefit by grouping items based on features in common. One well known example of such machine learning method is CART, and its dominant advantage is its simplicity as a classifier(Breiman, 2001, Goldstein et al., 2011, Papana and Ishwaran, 2006). If trained on high quality data, it can result in great accuracy when predicting test data. The principle is that every question is contained in a node, and every internal node points to one child node for each possible answer to its question. That creates a hierarchy that is encoded as a tree, data tree object. An object can be assigned to a class by following the path from the topmost node, the root, down to a leaf, according to the answers that apply to the item under consideration. An item is assigned to the class that is related to the leaf. Decision trees are sometimes more interpretable than other alternatives as, if one desires to, one can follow the exact path that leads to a decision (yet with the danger of overfitting). All sorts of features can be handled, including both categorical missing values and real-valued missing values. In this particular study, several promoter descriptive features were considered including the frequency of a motif, its position and the motif sequence. Furthermore decision trees can handle multiclass problems. Moreover, measures were established to account for impurity of a group (showing us how well a class can be separated), and these include entropy and Gini index. Specifically, given a measure of impurity I, we chose a question that reduced the weighted average of the impurity of the resulting children nodes. If I is the entropy function, then the difference between the entropy of the distribution of the classes in the parent node and this weighted average of the children's entropy is known as the information gain. Although there are many benefits, they come with dangers of over-fitting the training set.

Decision trees with all their listed benefits can be used in even more powerful manner, that is, in the form of a random forest ensemble learning (RF)(Breiman, 2001).RF trees differ in principle from CART as they are grown nondeterministically using a two stage randomization procedure (cite Ishwaran). Such ensembles are the best type of a classifier for many problems in molecular biology and other domains of research. In random forest, there is a randomized tree building algorithm. Through maintaining a collection of good hypotheses, rather than committing to a single tree, the chances are reduced that a new incoming instance will be misclassified. It will unlikely be misclassified as many trees are used for training. RF trains an ensemble of individual decision trees based on samples, their class designation and variables. Every tree in the forest is built using a random subset of samples and variables. A decrease in Gini impurity is linked to an increase in the amount of order in the sample classes introduced by a split in the decision tree. Every tree gives a vote for the sample after which the majority vote determines the class of the sample. The output consists of variable importance measures yet additionally one can determine the proximity between samples. There are several examples where the theory behind the RF was successfully implemented to solve a range of problems. For example, Allen et al. used decision trees within the JIGSAW (Allen et al., 2006) system to combine evidence from many different gene finding methods, resulting in an integrated method. Several studies by Ishwaran et al., aimed at explaining diseases classification given a series of SNPs (Papana and Ishwaran, 2006). It became one of the best available ways to find genes in the human genome and the genomes of other species. There are other hallmarks of this unique classifier such as, for example, the ability to uncover interactions using both reconscution and inherent pairwise variable importance measure. These conditional relationships can, in principle, be discovered within the data with RF as these are implicitly taken into account by the algorithm during the design of the classification model. Interactions between variables will often go hand in hand with conditional dependencies between the variables, for example variable B contributes to classification given that variable A is present above B in the tree. Conditional relations will be evaluated throughout the thesis. RF effectively handles the curse of dimensionality (Friedman, 1997). Let d represent the data dimension and n represent the sample same size. The d in terms of gene expression and SNP data usually spans between d being 102 and 106 The problem is incurred as when the dimensionality d increases, the volume of the space increases so fast that the available data becomes sparse. For one to obtain a statistically sound and reliable result like to estimate multivariate functions with the same accuracy as functions in low dimensions, we should establish that the sample size n will grow exponentially with d. The output can be viewed at different levels, the global and the local. The global level is the assessment of the entire problem; it captures the classification impact of variables on all samples. The local on the other hand is the estimate of the importance of a variable for the classification of a single sample. Local importance may uncover the specific variable importance patterns within groups of samples that are of interest for particular genes not the global problem. Particularly in recent years, the RF algorithm became popular, for example, in the genome wide association studies (GWAS), popular for pattern recognition in omics-scale data, and all this as a result of its two most important characteristics: the high accuracy and variable importance information (not to mention the speed of computation). Apart from RF per se several adaptations of RF have been proposed like the random survival forest proposed by (cite Ishwaran) which is highly effective for both prediction and variable selection in high dimensional problems. It can model non linear effects and interactions, can be used for event specific selection of risk factors and is a method free of model assumptions (Ishwaran, 2014).

In the upcoming paragraphs the results of the successful analyses using RF will be discussed and the evidence is provided. Nevertheless, it is worth mentioning the selection process that determined the features that were used. The selection of features relied upon the present knowledge of the architecture of plant gene expression. The analyses demonstrate the likely effect of frequency and position. As it is likely that the arrangement and the type of binding sites describes the architecture of the promoter it was interesting to integrate such features. Both supervised and unsupervised analysis was carried out to uncover the heuristic rules within the circadian system. Although there is not a set of golden rules, there are some heuristic rules that do operate in other systems as studies by Cox et al., on synthetic libraries revealed as the fact that repression dominates activation giving rise to asymmetric logic, the importance of location, and the importance of proximity in inducing AND logic. The summary of the rules of operation is portrayed by Figure 4.1a through 1d which reveals the logic behind the design of the studies. Apart from logic example there are instances from specific genes like GIGANTEA. Using phylogenetic shadowing three evolutionarily conserved regions within the promoter were identified proving that this combination is sufficient to confer a similar transcriptional pattern as the full-length promoter (Markus, 2014). The dissection of those regions showed that one subfragment contributes light inducibility, while another elicits a diurnal response (Markus, 2014). Altogether it was shown that 3 ABREL motifs together with 3 EEs give rise to diurnal expression. This pattern is characteristic of other genes expressed in the evening. Of great interest is the frequency that was detected specifically two types of conserved cis elements, each occurred three times within the promoter region. Mutation of all six motifs completely aboilished any diurnal rhythm in expression. Even better was the observation that particular distances between these elements are strongly statistically overrepresented. The authors speculate that this is the case in order to allow interactions between protein complexes bound to these motifs. More such examples are being elucidated with the well known one being described by (Li et al., 2011b) where there is a clear distinction between activation and repression achieved through action of different elements ELF4, LHY and CCA1 bind to EEs acting as repressors of light induction that is conferred through FHY1-FAR1 binding sites: (CACCGCG) and ACGT-containing elements (Li et al., 2011b). The specific rules are depicted on Figure 4.1.

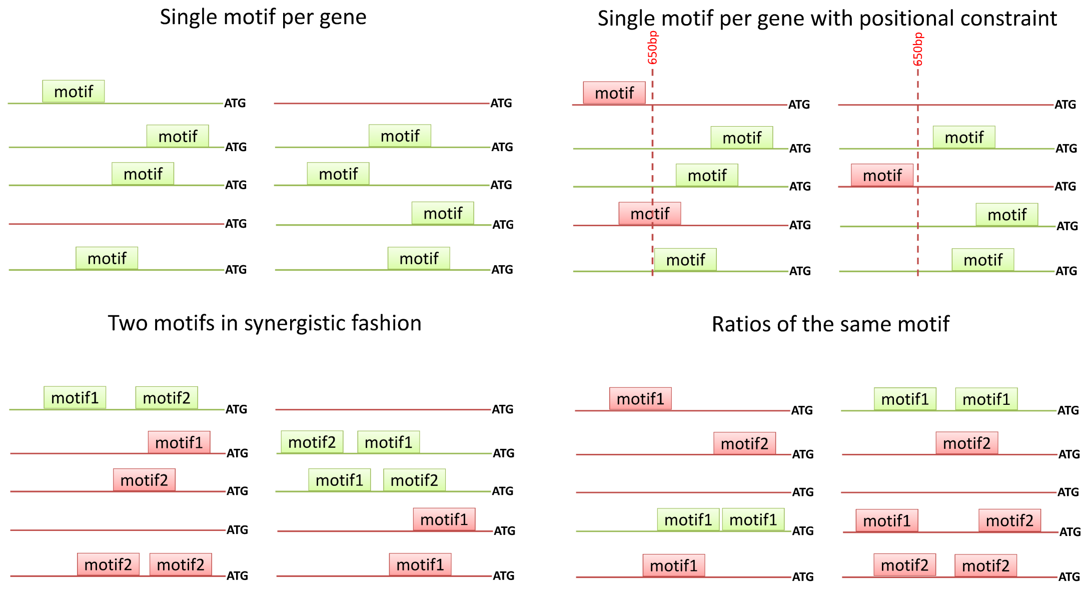

Single motif per gene
Single motif per gene with positional constraint
650bp
Two motifs in synergistic fashion
Ratios of the same motif
motif
motif1
motif2
ATG

FIGURE 4.1: The combinatorial logic behind the search space is portayed in four scenarios: top left quadrant single motif per gene; top right quadrant single motif per gene with positional constraint; bottom left quadrant two motifs in synergistic fashion; bottom right quadrant ratios of the same motif. The red colour portrays absence whereas green portrays presence of a motif in a promoter with respect to ATG.

The search for the hallmarks of circadian phase classes evolved as a result of thorough experimentation. The presented methods yielded accumulating results and formed the foundation for further experimentation, nevertheless not all results were used. Single trees were grown on the previously detected groupings presented in Chapter 3 (the 4 modules). The results section of this chapter contains solely the classifications that generated results of interest. The remainder of the paper is organized as follows. In section 4.2.1 the outcome of random forest classification is described. In section 4.2.2 random forest on alternative input matrices, specifically on interactions is presented. In section 4.2.3 game theory concepts are extrapolated and the analysis is from the level of a single gene, rather than the model. In section 4.2.4 the synthetic promoters devised based on previous sections are presented whereas section 4.2.5 concludes with the results of Dyck path representation analysis and its derivatives which constitutes alternative approach for such cis motif analysis.

4.2 RESULTS

4.2.1 IDENTIFICATION OF CIRCADIAN PROMOTER ELEMENT COMBINATIONS USING RANDOM FOREST

We sought to make use of the novel groupings presented in Chapter 3 to extend our characterization of the circadian system by looking for *cis* elements which could explain patterns of expression in the Arabidopsis data. We reasoned that the phase modules of circadian genes classified by projection onto inherent trends within the data would provide more realistic phase groupings than those used previously which relied upon artificially imposed phase group boundaries. For this we also applied a novel approach. Previous attempts to identify circadian *cis* elements have looked for enrichment of single elements within phase groupings (Bussemaker et al., 2001). Building on recent research suggesting that multiple elements, in fact, act co-ordinately to generate a specific circadian pattern, we used a method which would identify such coordinately-acting groups (Beer and Tavazoie, 2004). For effective utilisation of the large number of *cis* elements likely to be involved in such multivariate responses the random forest (RF) methodology was used to predict important motifs. One kilobase pair length of promoter sequence was analysed for each gene. Random Forest outputs creates ensemble of trees like one mentioned in Figure 4.2. This particular example is a frequency motif tree. The most important split in this unpruned tree is AATATC, a split based on precence/absence. The next informative feature is having/not having GGCCCA in a frequency of two and having in frequency of 3, ATTTTTG. Due to its extensive size the tress were pruned. All possible 5-8-mer sequences were considered. Of these the sequences showing the highest enrichment as isolated elements in the circadian dataset were chosen for the combinatorial analysis.

RF seeks to assign new samples to specific groups or classes based on features in common with other members of that class, in this case, *cis* elements. It uses a decision tree system of classifying, that is, it asks whether one feature at a time is present or not thus producing two branches. These branches then branch further as additional features are considered. Ultimately, these decisions about the features of a sample (branches) lead to its assignment to a specific class. In a RF, an ensemble of decision trees is created. Each individual tree is grown from a randomly sampled subspace of input features (*cis* elements from among the 21 highest-enriched individual elements) and final classification is made by combining results from trees via voting. It is a machine learning approach which, makes use of subsets of data to capture these features of interest. The learning element of this approach comes from the way in which these decision trees are initially created using two thirds the dataset. The decision tree is then re-created using the remaining subset of the data to assess whether the same classifiers can correctly assign the members of this subset. If not, a new decision tree is created. The RF was applied subsequently due to its high suitability, specifically as it is hypothesis free and unconstrained by a priori assumptions. The fundamental element to the analysis of predictions of class is the effect upon bias and variance. Bias is large if the learning method produces classifiers that are consistently wrong. It relates to the ability of a given model function to approximate the data, and so high bias is related to under-fitting. Variance is the variation of the prediction of learned classifiers, it pertains to the stability of the model. High variance learning methods are prone to overfitting the training data. The decomposition for prediction error loss having a continuous outcome is equal to noise plus bias and variance. Whereas the bias is referred to as the systematic difference between the prediction and the target, the variance is defined as the measure of randomness of prediction. When aiming for the unbiased classifier we are looking at low variance. For every classification a bootstrap sample of the data was selected. In the search the optimal splits were detected. This was repeated until an unpruned, large and unlimited tree was made. The data not being part of the bootstrap sample was run down and tree and the error rate was noted together with variable importance measure. The input data matrices are deposited in the supplementary data. Cross validation is inherent to the Random Forest methodology. Specifically 2/3 cross validation procedure was used in this analysis. Nevertheless, there is no need for cross-validation to get an unbiased estimate of the test set error as it is estimated internally. Trees are constructed using a different bootstrap samples from the original data. Benchmarking studies have shown that RF displays increased performance when compared with individual trees. The analysis showed that expression patterns of

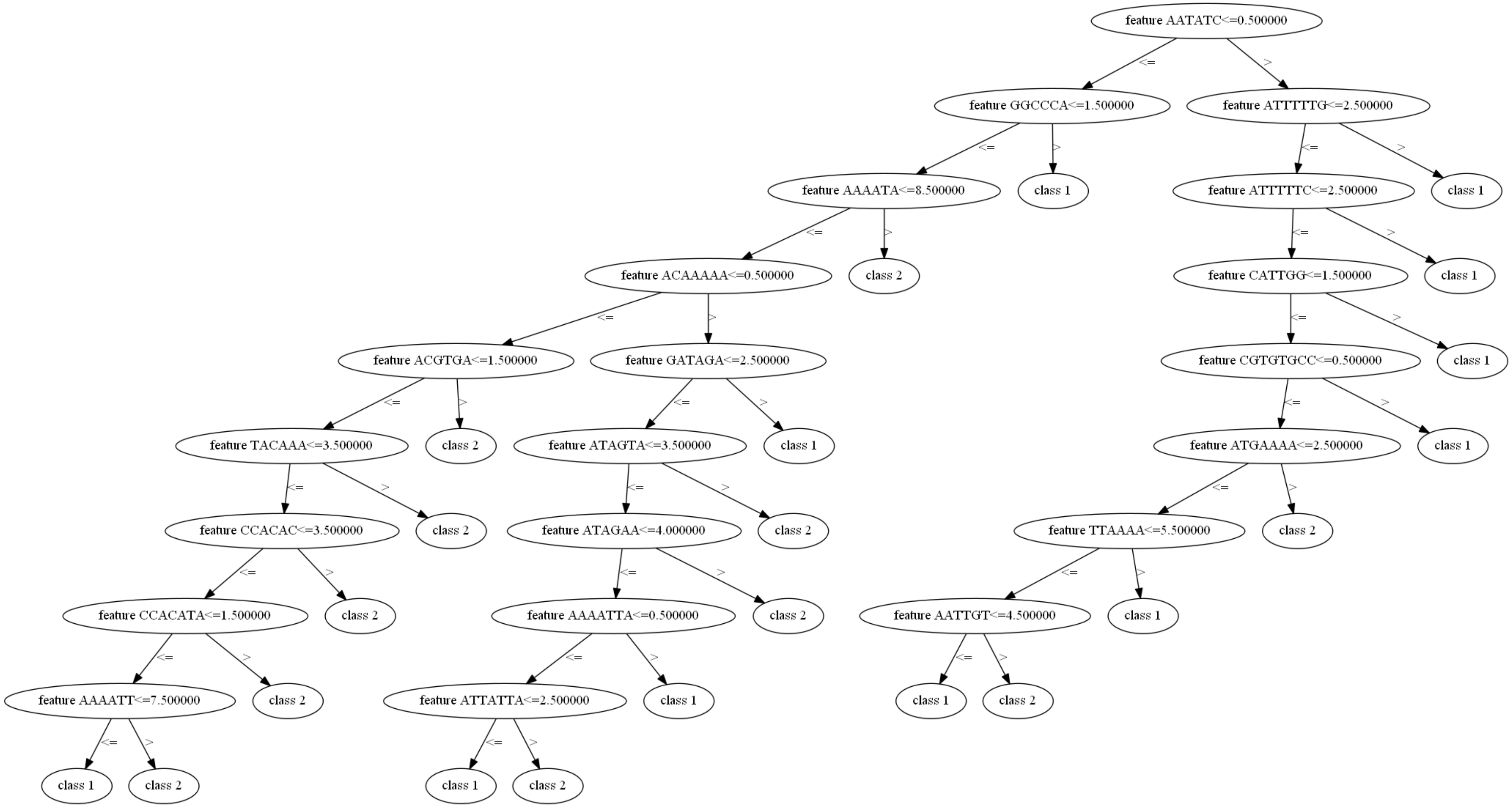


FIGURE 4.2: Output tree from Random Forest which account for both motifs and their frequencies. In this particular tree the most importance split is that of having not having AATATC. The 'greater than' annotation at each split indiciates more than x frequency of that particular feature whereas 'less than or equal to' annotation indicates having less of a given motif.

circadian genes can be predicted relying on combinations of upstream *cis* elements. Figure 4.3 shows motifs ordered by variable importance measure (VIMP, the relative contribution of that variable or motif to the classification of the genes) for each of the 4 phase modules. To evaluate the presented cis elements we randomly permuted the exact locations of the binding sites for the tested element within each promoter within a given phase module (at the same time) preserving the number of binding sites per module and repeated the computation for randomly placed sites. Each permutation was performed 1000 times. As we are seeking to define a particular multivariate response defining a phase module so to say it is impossible to test on a benchmark circadian set in the latter case yet the results provided in Figure 4.4 do demonstrate the significant difference in the VIMP score between the observed and random elements. The results here are combinatorial in that the data represent a group of *cis* elements which, in the context of other elements, are predicted to faithfully determine the phase of the genes within each phase module. Notably, several of the detected *cis* sequences giving the highest VIMP in combinatorial analysis form part of elements previously described in literature; for example. AATATC, part of the Evening Element, involved in regulation of a number of circadian genes in Arabidopsis (Harmer, 2000, Hsu et al., 2013); GATAA, part of the I-box, involved in response to light (Giuliano et al., 1988); and CAAAA, part of the *CAB2* DET1-associated factor 1 binding site (CDA-1) in the dark response element, involved in response to darkness (Maxwell et al., 2003). It is important to note that it can be either the presence or the absence of these *cis* elements which can be indicated. In order to better illustrate how these elements may contribute to the expression pattern of individual genes within a phase module, Figure 4.5 demonstrates a prototypic pattern per class using example genes (*CCA1, TOC1, PRR7* and *CYP96A4*). The promoters of these genes were searched for the elements identified as being important in determining inclusion in the respective phase module. The contribution of each of these elements was also calculated by determining the proportion of genes correctly assigned to the phase module by each element (a negative value indicating absence as being important). It can be seen that, in general, for each of these genes, there is a high occurrence of the *cis* elements whose presence is shown to be important in assigning genes to that phase module supporting the proposal that these *cis* element combinations could genuinely be important biologically.

The overall performance of random forest can more objectively be assessed using Receiver Operating Characteristic (ROC) curve. The ROC curve can be thought of a the quantitative summary of the power of the employed test and ranges between 0 and 1 with 1 indicating high power. A ROC curve shows the false positive rate, number of false positive predictions, as a proportion of the total number of negative predictions along the x-axis and the true positive rate, number of correctly predicted positive predictions, as a

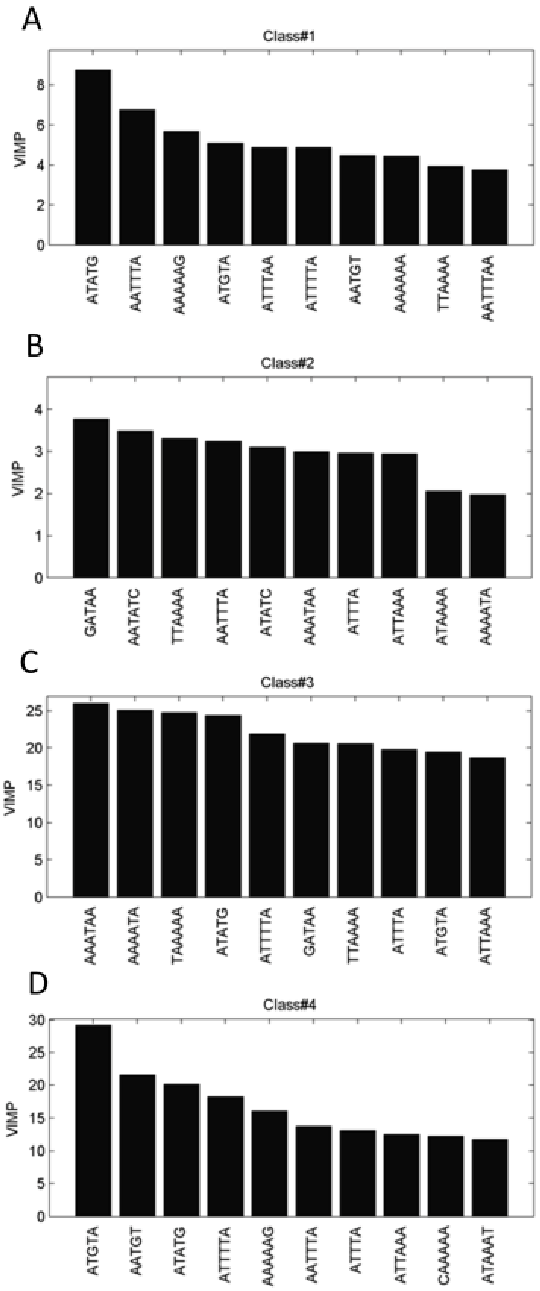


Figure 4.3: *Cis* elements defining phase of expression for circadian genes in Arabidopsis. Circadian genes identified in Arabidopsis circadian timecourse data were divided into four phase modules based on projection onto circadian ICA components. Random Forest was used to identify *cis* elements which collectively act as classifiers for each of the four modules. Classes 1 and 2 (A and B) correspond to subjective dawn and subjective dusk respectively (2nd component); classes 3 and 4 (C and D) correspond to the middle of the subjective day and night respectively (1st component). Motifs were selected and ordered through their variable importance (VIMP).

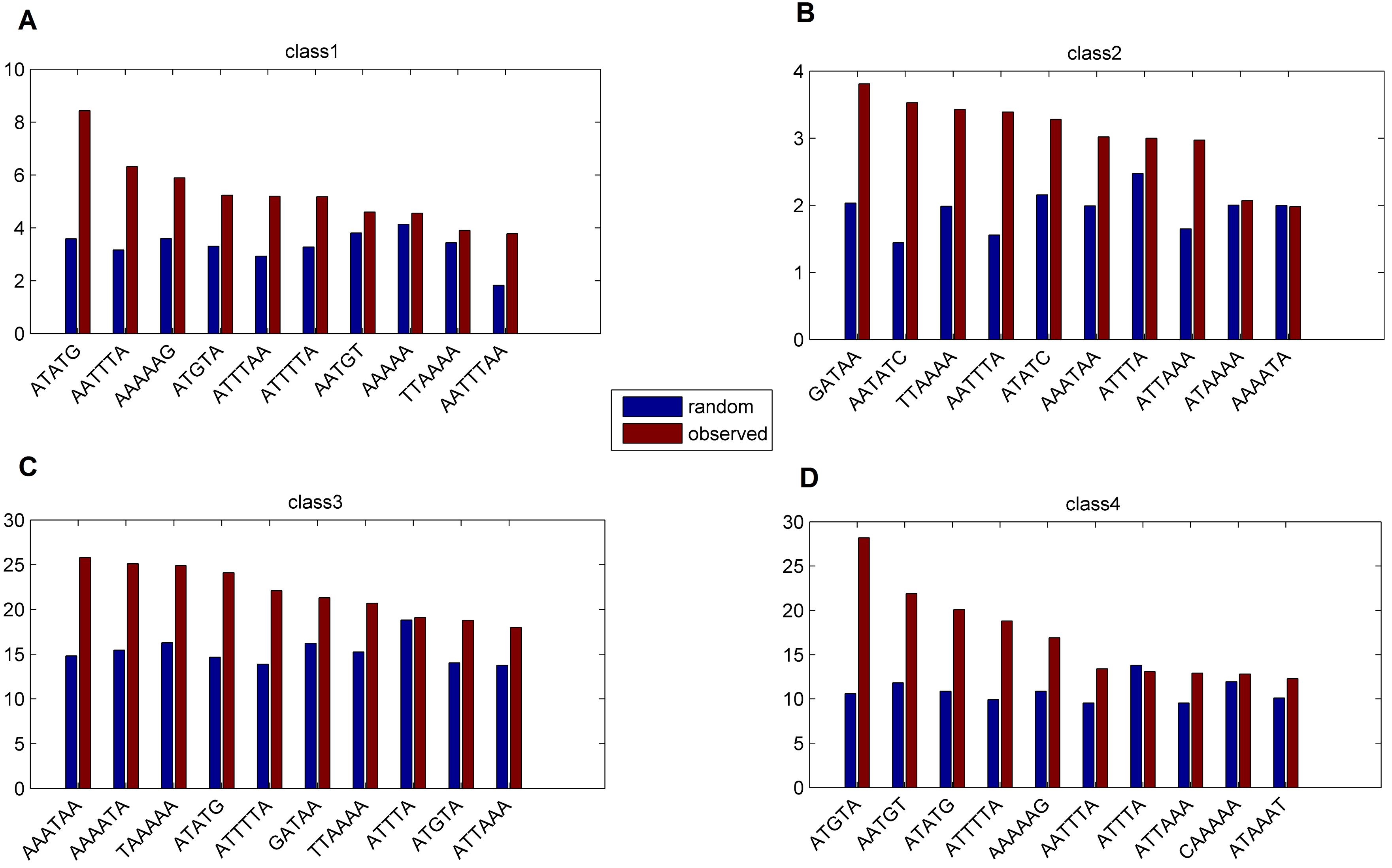


FIGURE 4.4: *Cis* permutation. To evaluate the presented cis elements we randomly permuted the locations of the binding sites for the tested factor within module (at the same time) preserving the number of binding sites per module and repeated the computation for randomly placed sites. Each permutation was performed 1000 times. Figure 4.4 demonstrates the significant difference in the VIMP score between the observed and random elements.

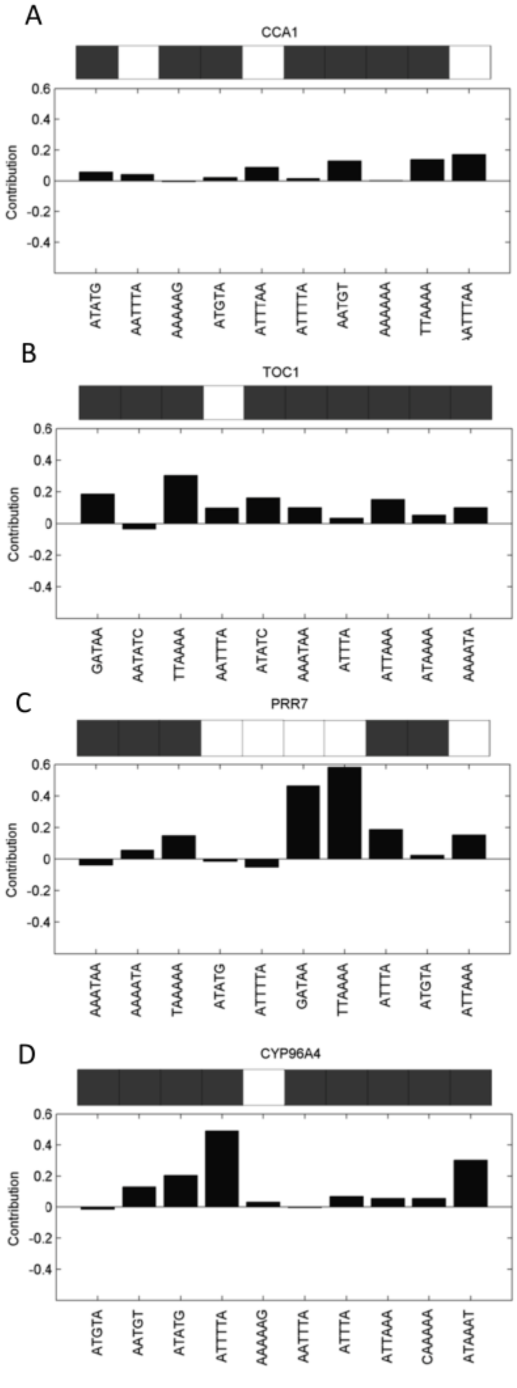


FIGURE 4.5: Contributions (RF votes data) of individual features are presented for four correctly predicted prototype genes, representative of the four distinct phase modules or classes. RF voting calculates the association of either the presence and the absence of a feature with a particular class as indicated by a positive or negative contribution. Presence or absence of a motif in the prototype genes is represented above each graph with filled boxes representing presence and open boxes representing absence. (A) *CCA1* (Circadian clock associated 1; peak phase subjective dawn), (B) *TOC1* (Timing of CAB 1; peak phase subjective dusk), (C) *PRR7* (Pseudo response regulator 7; peak phase middle of subjective day), (D) *CYP96A4* (cytochrome P450, family 96, subfamily. A, polypeptide 4; peak phase middle of subjective night).

proportion of the total number of positive predictions along the y-axis. A perfect ROC curve would be a horizontal line y=1. A ROC curve expected from random guessing is the diagonal y=x line. A common metric for assessment of performance is the area under the curve (AUC) value. A perfect AUC is 1 and a value of 0.5 is expected from random guessing. On a model level, results can have precision and recall. Precision is the success rate in predicting the correct class as opposed to the incorrect class, while recall is the success rate in predicting the correct class as opposed to not assigning to any class. Figure 4.6 portrays ROC curves for the ability of the identified *cis* elements to correctly predict the four phase modules. The AUC values for the four classes are 0.9420, 0.9429, 0.9567 and 0.9449. A very high area under the curve in all cases indicates that the identified combinations of promoter elements can be considered meaningful with a high degree of confidence.

What was most interesting about the groups of elements associated with each phase module was the way that the individual elements making up the group changed progressively from one phase module to the next. When the elements common to more than one classifier were aligned, along with the previously-highlighted elements of interest, in a simple presence/absence table, a pattern emerged (Table 4.1). The table was manually selected for.In order to test out hypothesis that progressive combinations of individual cis elements acting in concert could be responsible for the range of possible phases of Arabidopsis circadian output genes, we compared the elements associated with each phase module. We did this by aligning the elements common to more than one phase classifier, along with the previously-highlighted elements of interest, in a simple presence/absence table. We observed that individual elements making up the group associated with each phase did, indeed, change progressively from one phase module to the next (Table 4.1). Although the specific combination of elements was unique to each phase module, each module showed some overlapping elements with the next in sequence. Furthermore, the elements showed a progressive pattern of change from one phase module to the next. Moving from one phase to the next through the day sees the addition and/or removal of elements to the group identified as determining the previous phase. This strongly supports our proposal that the additive effect of a combination of elements acts to determine the specific final phase. For example, although the midnight, dawn-phased and noon phased groups of genes all share a common core of four elements, the table suggests that removal of the CAAAA element from the dawn-phase element combination shifts the timing of genes from the dawn-phased group of genes to the earlier midnight-phased group of genes. Conversely, addition of the GATA element, GATAA, and removal of AAAAG, AATGT and AATTTA to the dawn-phased element combination shifts the timing of genes from the dawn-phased group of genes to the later noon-phased group of genes. In order to confirm the significance of the element groups

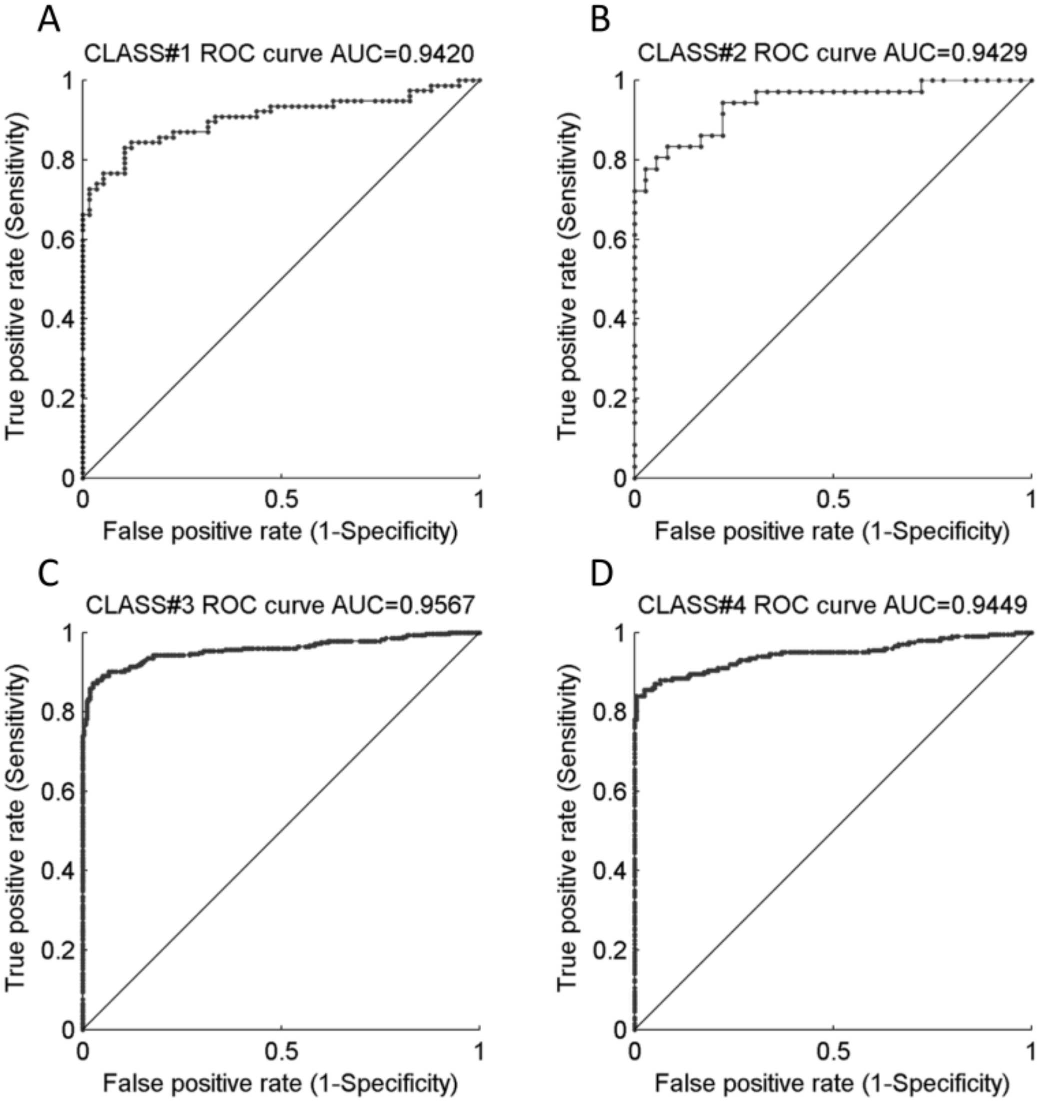


FIGURE 4.6: Assessment of *cis* element classification performance. Receiver operating characteristic (ROC) curves represent the power of the feature combination classifiers identified by Random Forest to correctly assign genes to our four phase modules. Classes 1 and 2 (A and B) correspond to subjective dawn and subjective dusk-phased genes respectively, whereas classes 3 and 4 (C and D) correspond to the middle of the subjective day and night respectively. The area under the curve (AUC) scores are also represented above each graph.

described here, we identified the four groups of our circadian genes containing the element groups described in Table 4.1 and analysed their mean phase. The mean position of each group of genes when projected onto the two key circadian eigentrends defined by ICA is represented in Figure 4.7. For each group of genes the mean phase is very close to the component axis representing the expected phase based on their element combinations, while the mean magnitude of dot product in each group is above 0.8, the cut-off used in our definition of circadian genes. The tick and cross presentation is chosen over a quantitative measure as the focus is on combinations of elements. The mean phases radial plot (Figure 4.7) specficially demonstrates that point. Additionally it is worth mentioning that there are other important elements that are not depicted in the table. It should be re-emphasised that the reason they are not in the table is just that they are omitted to allow the progressive aspects of the cis element patterns to be visualised more clearly.

It was observed that, although the specific combination of elements was unique to each phase module, each module showed some overlapping elements with the next in sequence. Furthermore, the elements showed a progressive pattern of change from one phase module to the next. This strongly supports the proposal that a combination of elements acts to determine the specific final phase with removal of certain elements and addition of others allowing the phase of a gene to be earlier or later as required.

4.2.2 RANDOM FOREST ON INTERACTIONS

The results presented in the first section led to some further questions being raised about the matrices being classified and the tuning properties. Next the aim became testing random forest on alternative matrices using interaction data as input. That called for several changes in the tuning parameters. As the figures and data presented below will explain, that led to higher order insights upon the drivers of the phases. The below section throughly explains one such classification that gave highly reproducible and significant results.

Several transcription factors often work together, to enable the cells to respond to various signals. The detection of combinatorial regulation by multiple transcription factors, however, is often overlooked. One could go back to the lac operon for proofs yet that is not the aim here. When combinations of up to k motifs are considered, the number of tested combinations increases exponentially with k. Strictly mathematically, because of the exponential growth in the number of tests, the discovery of combinations of greater arity (encompassing many motifs) motif combinations is difficult in principle, even with more sensitive corrections like the recently proposed limitless arity multiple-testing procedure (Terada et al., 2013). Here such approach is represented on Figure 4.8 where the input to RF was a set of motif pairs.

The performance was assessed in Figure 4.9 by ROC curve and precision recall curve which shows great performance, that is above 95%. The ROC curve describes the overall performance of the classifier over positives and negatives, whereas the precision-recall curve is of interest because it does not include the true negatives, and hence the emphasis is on how well the true positives are reconstructed. Figure 4.10 demonstrates a prototypic pattern per class using example genes (CCA1 and LHY, TOC1 and ELF4,

	AATATC	ATATC	GATAA	ATTTA(A)	ATATG	ATGTA	ATTTTA	AAAAG	AATGT	AATTTA	CAAAA
Midnight	X	X	X	✓	✓	✓	✓	✓	✓	✓	✓
Dawn	X	X	X	✓	✓	✓	✓	✓	✓	✓	X
Noon	X	X	✓	✓	✓	✓	✓	X	X	X	X
Dusk	✓	✓	✓	✓	X	X	X	X	X	✓	X
Midnight	X	X	X	✓	✓	✓	✓	✓	✓	✓	✓
Dawn	X	X	X	✓	✓	✓	✓	✓	✓	✓	X
Noon	X	X	✓	✓	✓	✓	✓	X	X	X	X
Dusk	✓	✓	✓	✓	X	X	X	X	X	✓	X

TABLE 4.1: Comparison of *cis* elements classifying phase of expression. *Cis* elements contributing to more than one phase of expression or forming part of previously recognised light or circadian elements are shown for each of the four phase modules based on projection onto circadian ICA components. Phases are double plotted to highlight the progressive cyclic pattern of element contribution.

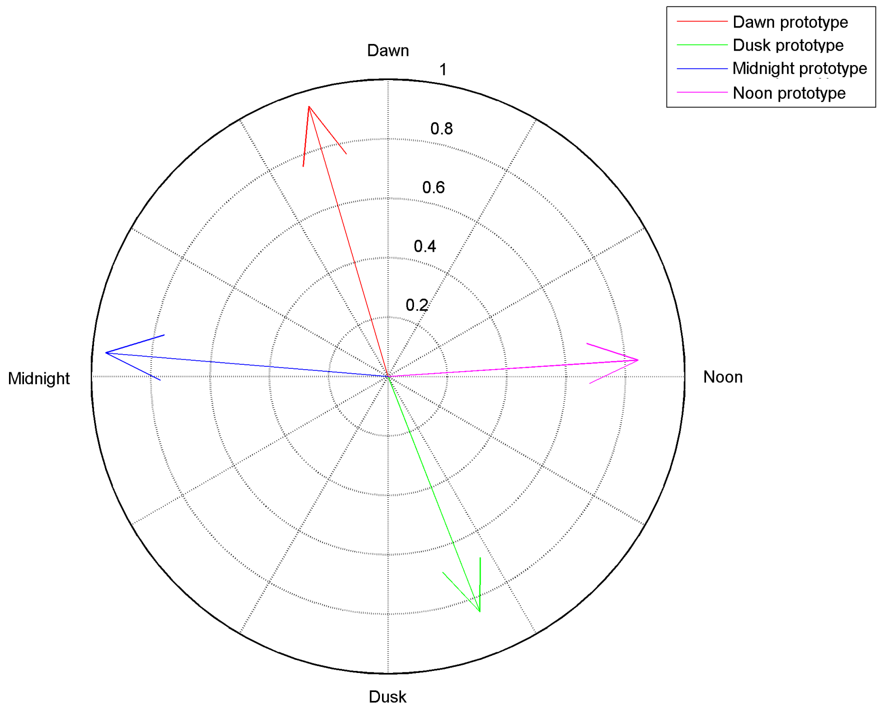


Figure 4.7: Mean phase of genes possessing progressive *cis* element combinations identified by Random Forest. The tips of the arrows represent the mean dot product projection of each of the four groups of genes possessing the distinct cis element combinations identified in Table 4.1 onto the first (x-axis) and second (y-axis) circadian ICA components. The radial axis represents the dot product value of the projection.

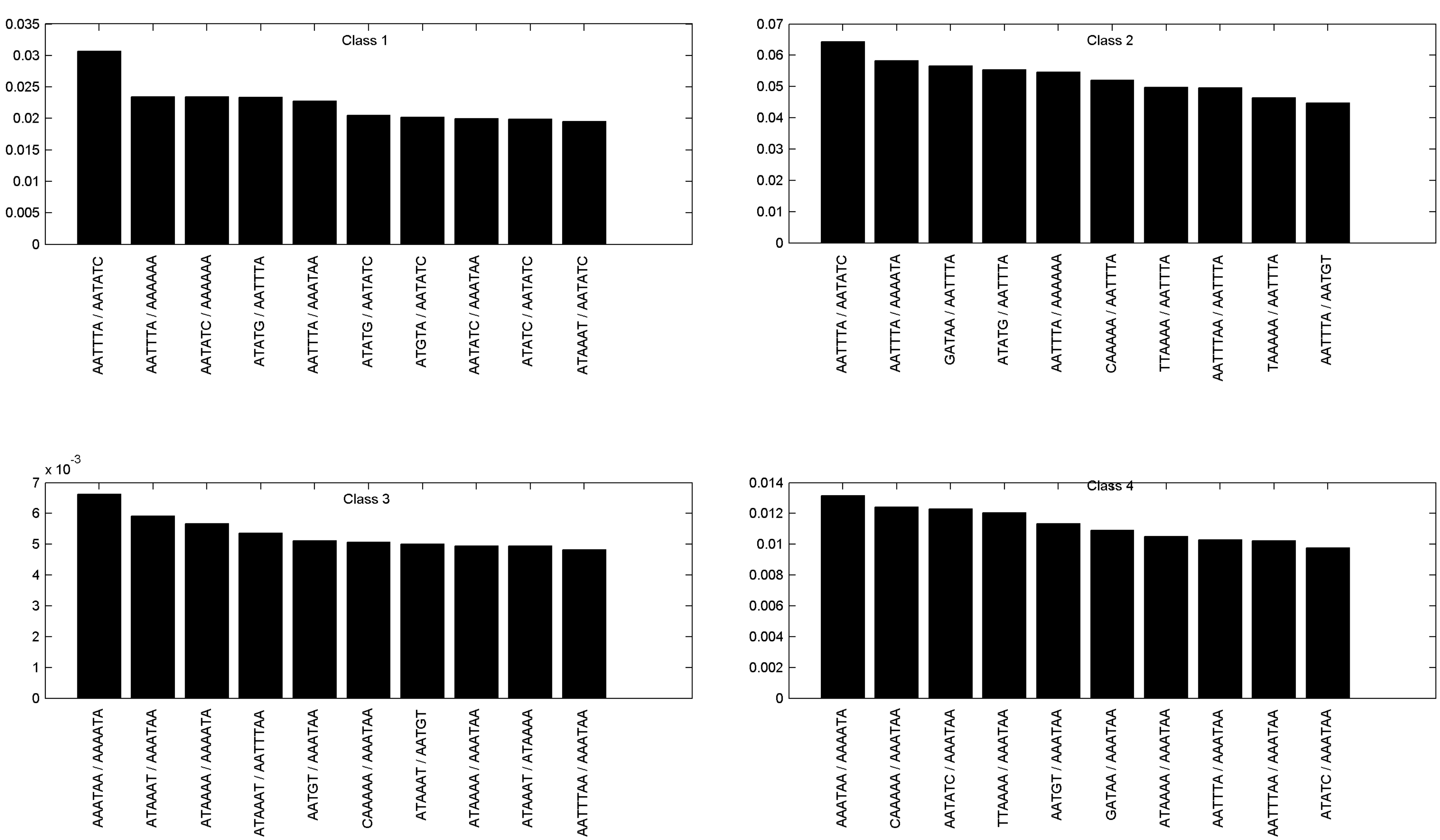


FIGURE 4.8: The most important motif pairs in terms of their contribution to the 4 classes of circadian regulation. *Cis* elements defining phase of expression for circadian genes in Arabidopsis. Circadian genes identified in Arabidopsis circadian timecourse data (GSE8365 dataset) were divided into four phase modules based on projection onto the two circadian eigentrends. Random Forest was used to identify *cis* elements which collectively act as classifiers for each of the four modules. Classes 1 and 2 (A and B) correspond to subjective dawn and subjective dusk-phased genes respectively (2nd component), whereas classes 3 and 4 (C and D) correspond to the middle of the subjective day and night respectively (1st component). The most important motif pairs which distinguish members of individual classes were selected and ordered by their contributions. The scale represents change in prediction of a given class for a given pair of motifs. The bigger the value the greater the importance of two motifs acting in cooperation.

PRR7 and CYP96A4). It is interesting to compare this these prototypic genes using different input sets.

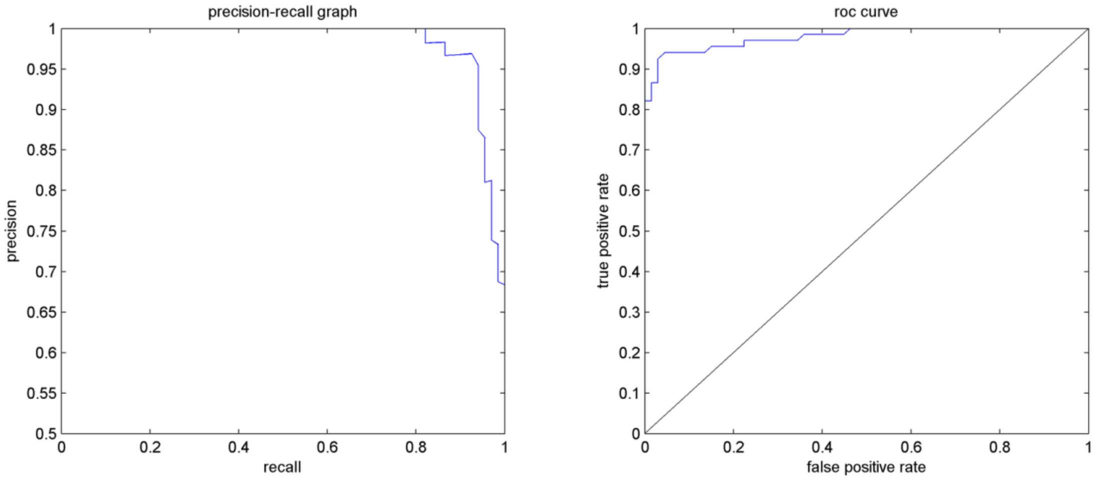


Figure 4.9: Precision Recall and ROC curve is portrayed for one of the classification results showing the high information content of the *cis* elements given the solid class prediction in both Figures.

Performance at this stage is indicative of the entire model. Unfortunately, in the specific instance, the variable importance measure answers half of the posed question as the second half is determined by the motifs per se. One way to establish a cutoff on a variable importance plot would be through the so called elbow cutoff where there is a significant fall in the contribution of a given motif for example. The main tuning parameters that fundamentally affect trees include: the mtry which is number of randomly selected variables used in each split, the number of trees and the tree size. If mtry which is the size of the subset is small, that will result in trees with lower correlation, effectively potential for variance reduction. When the mtry is equal to the number of predictors, RF reduced to bagging which is a method of combining multiple predictors. In principle, larger values of mtry lead to fewer variables brought into the tree, resulting in sparse solution. Essentially, mtry can be thought of as controlling the degrees of freedom of the model, with the higher mtry the fewer df used (Goldstein et al., 2011). In the results the selected mtry was defined by the square root of the total number of predictor variables. The assumption here is that the greater the number the trees (ntrees), the better. Basically the only limitation here is computation. The dominant reason behind selection of the tree size is to reduce the potential for overfitting. It can be controlled through the number of splits. Pruning in principle increases stability, lowering the variance at the cost of higher bias. The class weights were proportionally assigned.

The performance was assessed using precision recall curves and ROC curves, and supplemented by CROC curves (Swamidass et al., 2010). As ROC curves have some

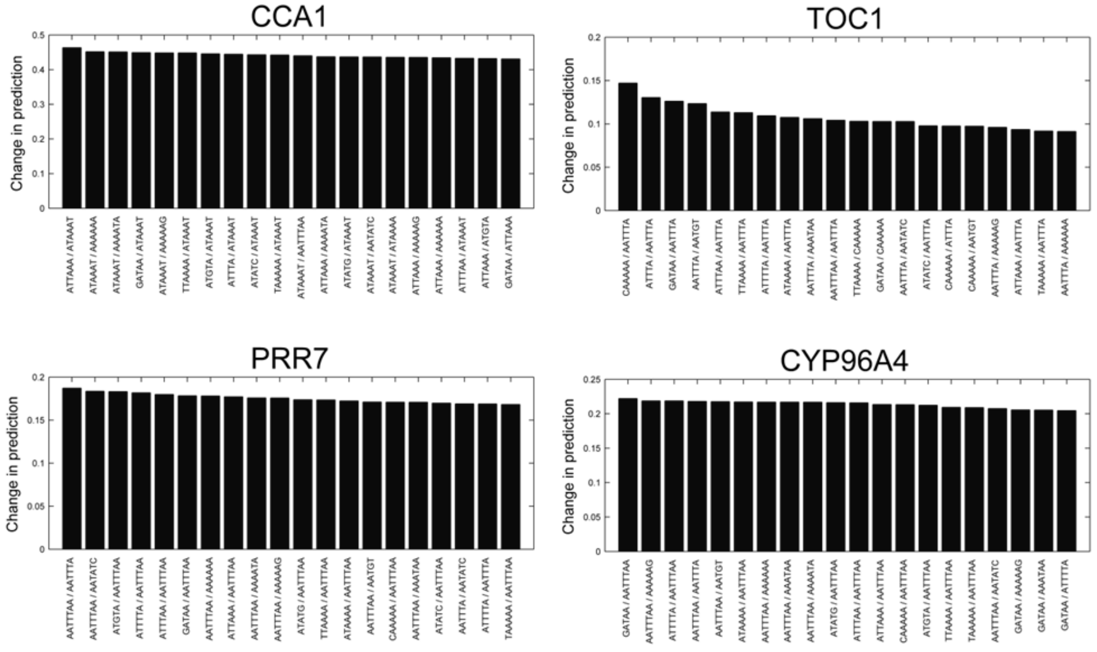


FIGURE 4.10: Application of *cis* element contributions predicted by Random Forest model. Contributions (RF votes data) of individual features are presented for four correctly predicted prototype genes, representative of the four distinct phase modules or classes. RF voting calculates the association of either the presence and the absence of a feature with a particular class as indicated by a positive or negative contribution. (A) *CCA1* (Circadian clock associated 1; peak phase subjective dawn), (B) *TOC1* (Timing of CAB 1; peak phase subjective dusk), (C) *PRR7* (Pseudo response regulator 7; peak phase middle of subjective day), (D) *CYP96A4* (cytochrome P450, family 96, subfamily. A, polypeptide 4; peak phase MIDDLE OF SUBJECTIVE NIGHT)

limitations. For example summarizing, overall, the possible threshold may indeed be misleading, the extreme left and right sides may be of no relevance regarding the problem at hand. To supplement these results every classification came with a precision recall curve and a CROC curve that essentially focuses on the interest area for the classification. CROC curves as proposed by Swamidass et al., address the early recognition problem and are able to decipher if 0.5 AUC models are true random classifiers. This comparison can be viewed in Figure 4.9. Figure 4.9 ROC and CROC curve comparison shows the applicability of CROC curve in such example where solely part of the ROC curve is of interest and could be magnified. Considering the confusion matrix, we are aiming at rather a level of true positives at the expense of false positives. We rather uncover a truly definite set of motifs committing to low false positive rate. The ideal would be to maintain the objectives; for example, having high sensitivity and high negative predictive value when trying to uncover early diagnosis in a medical case.

We inspected the change in contribution given pairs of elements. It confirms those elements mentioned as being of individual importance through VIMP yet further highlighting the intricacies in plant promoter design. For example, top dusk pair is GATAA

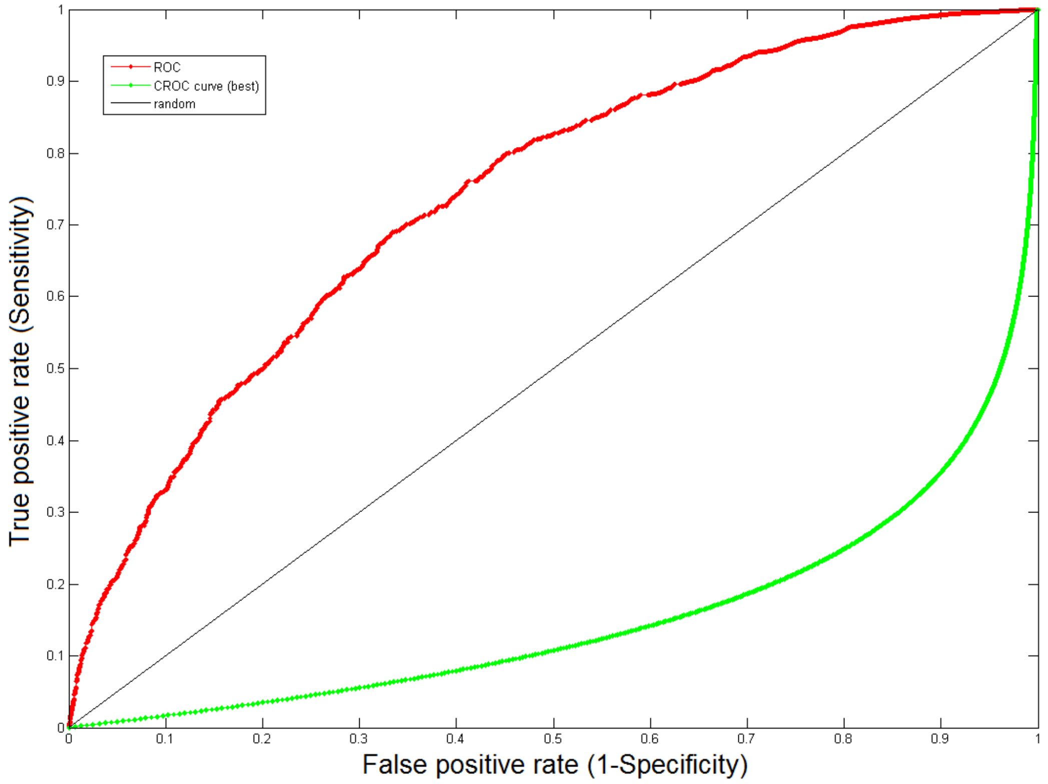


FIGURE 4.11: The distinction between the applicability of a ROC Curve and the CROC curve in terms of *cis* motif classification. The specific example is for a subgroup group of cis motifs in class 3.

and ATTTAA which together change the contribution by 12% (with 5 fold greater change when acting in concert). During the dawn phase having ATGTA and ATATG changes the contribution by 40% (with tenfold greater change when acting together than alone). For example AATATC and GATAA acting in concert are the strongest predictors for *ELF4* together affecting the entire contribution by 9% (6 fold increase through joint action). Such pairs are the first step to synthetic promoter design and a successful reconstruction of the natural circadian phases could suggest that the network structure of transcriptional circuits determines the fundamental timing of the outputs from the circadian system. The specific synthetic promoters are shown in Figure 4.18. Gratifyingly, as shown before we noted a significant enrichment of previously identified *cis* elements. It will be interesting to test these synthetic regulatory elements that do occur and that not occur in nature. To provide further validation of these predictions, we could employ a in vitro system of the clock to empirically test candidate elements in circadian transcriptional output assays (cell culture system that allows the monitoring of circadian transcriptional dynamics using a destabilized luciferase reporter driven by response elements). Furthermore if these in vitro validated elements do play a prominent role in vivo, the transcripts for these genes would likely oscillate in a circadian fashion. Using quantitative PCR assays, we could measure expression profiles from our predicted elements and evaluated their rhythmicity (ANOVA and curve fitting). Such results will further show how the amplitude information is encoded in specific residues adjacent to the core consensus element. Furthermore such studies will allow one to see the balance between trans activators and trans repressors in generating specific transcriptional output. Such validation of these *in silico* findings strategy *in vitro* and *in vivo* using real-time monitoring of transcriptional activity and fused luciferase reporter could lead to the identification of crucial elements that dictate rhythmicity and even more key clock genes given those combinations. It is the combination of system top down approaches and synthetic bottom up approaches here like design and testing of synthetic promoters in vivo that will provide altered plant oscillator. In random forests the *i*-th variables importance is obtained by randomly permuting its values in the out-of-bag examples and observing what this does to the error on the out-of-bag examples. More important variables will, on average, result in a bigger decrease in error and here the features are pairs Figures 4.10 through 4.12 demonstrate the top pairs. Figure 4.12 demonstrates the top pairs selected by VIMP per class. Circadian genes identified in Arabidopsis circadian timecourse data were divided into four phase modules based on projection onto the two circadian eigentrends. Random Forest was used to identify *cis* elements which collectively act as classifiers for each of the four modules. Classes 1 and 2 (A and B) correspond to subjective dawn and subjective dusk-phased genes respectively (2nd component), whereas classes 3 and 4 (C and D) correspond to the middle of the subjective day and night respectively (1st component). The most important motif pairs which distinguish members of individual classes were selected and ordered by their contributions. Figure 4.12 and 4.13 demonstrate the same principle when viewed from the perspective of 1 gene, here this being *CCA1*. The Appendix contains all the files where the prediction change is viewed when each of the motifs is considered separately versus jointly.

4.2.3 MOVING FROM THE MODEL SPACE TO THE GENE SPACE

Up to this point we generated several models that were trained on a sample of instances with features being the position, the frequency and the presence and absence of the motif per se with a class value. We assumed that the model is good at predicting the class value, but we want more insight into how it makes those predictions, insights into the black blocks. Another approach was employed to decompose model's prediction of individual contributions of each attribute. There are 3 dominant levels of explanation:

- The instance explanation; that is; the gene level where we consider one gene given the entire model Figure 4.13

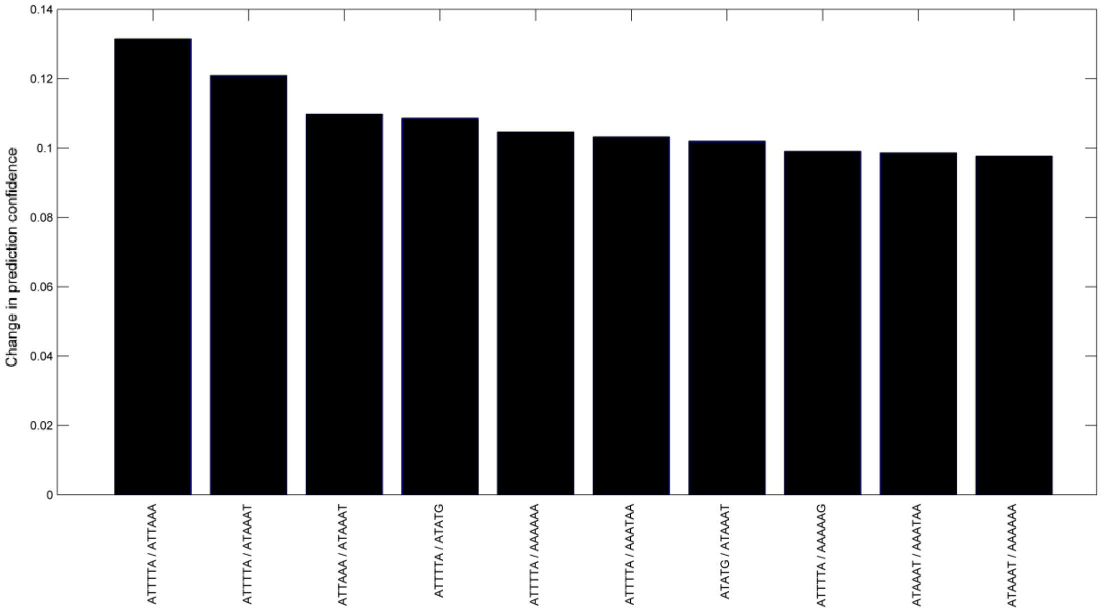


FIGURE 4.12: Shows the change in prediction confidence in terms of the most important pairs of motifs acting in concert. These are hallmarks of one class and were inspected in terms of their position. These are from a perspective of one gene.

- The model explanation: averages of explanations over several training instances, Figure 4.10

- The domain explanation, which may be similar to the model, yet that is not a given as the domain level uncovers the deterministic causal relationships, progression and dependencies.

The objective is to find out which features dominantly affect the thresholds for the change in prediction. The model here is the function of mapping instances into numerical values. In other words, the circadian model is a mapping from instance space to probabilities of the class values, thats is belongingness to a module. Every instance has a given value for each of its attributes. To find the effect that these values have on prediction one decomposes on individual attributes values and inspects the model's prediction. If the difference is large the attribute plays an important role. Predictions were evaluated using difference in probabilities. This could be done differently. The alternatives here would be assessment using information difference and the weight of evidence which are alternative methods to detect the major contributors. The probability difference is the difference in prediction of the model having the knowledge about the value of A_i and the outcome without it.

$$probDiff_i(y|x) = p(y|x)_i p(y|x\ A_i) \tag{4.1}$$

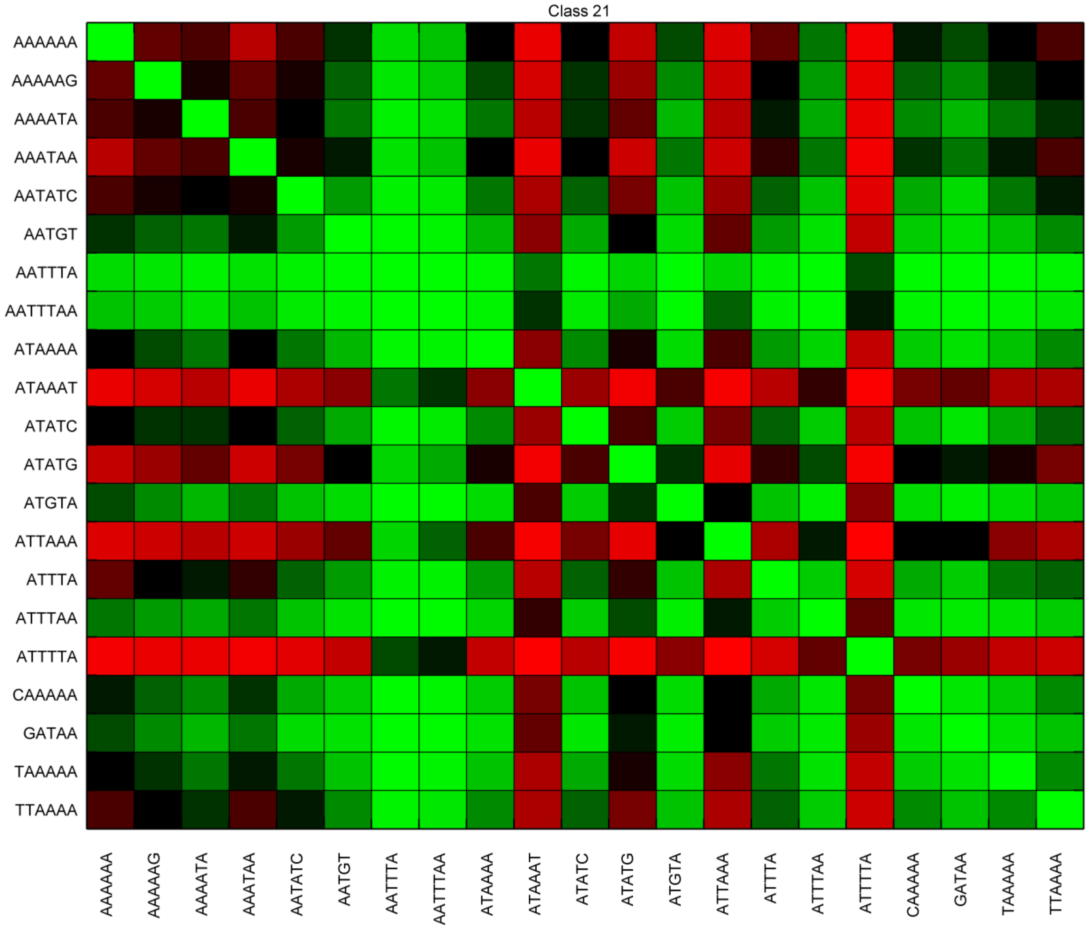


FIGURE 4.13: The interacting pairs of motifs within a specific promoter, here being *CCA1*, are portrayed. The heatmap portrays the probability of co-occurence of a set of motifs. The scale is between zero (light green) and one (red). Due to the nature of experimental design (maximum presence of each motif equals one) the diagonal effectively is zeroed out.

The points coming out of this analysis do shed some light into the organization of these four circadian classes. It seems that there is instance dependency as the genes do possess different combinations of the same motifs. For example the fact that GATA and EE elements appear in genes of various clusters, suggests that it is not a single element that determines timing of expression, contrary the series of elements changing in progression. Here the same pertains to the class which can be made of certain subgroups if there is dual functionality. The classes are made up of different elements. There are some strong conditional dependencies between individual attributes. Right away it is clear that some attributes do not give significant changes yet they are clearly distinct for the class, meaning that they act in concert. The focus on marginal predictions here is the key to the selection of individual motifs for design of synthetic promoters. Interpretation of such predictions is fairly intuitive, for example, using a Naïve Bayes classifier, where the assumption of conditional independence allows for direct inferences to be made. Yet we know that some features simply do not act alone, and, hence, CCA1

```
attcaaattacatgcatgcaactaagtagcaacaaagttgatatggccgagttggtctaaggcgccagattaaggttctggtccgaaag
ggcgtgggttcaaatcccactgtcaacattctcttttctcaaattaatattttctgcctcaatggATAAATAAATTAAAAATAAATA
AATTAAAAATAAATAAATTAAAAATAAATAAATTAAAAcggcccagtatcagttgtgtatcaccacgttatttcaaatttcaaactaag
ggataaagatgtcatttgacatatgagatatttttttgctccactgagatatttttctttgtcccaagataaaatatcTTAAAATTAAA
ATTAAAATTAAAAtttgcgcattaaaccaaaaagtgtcacgtgatatgtccccaaccactacgaattttaactcgaattttaaccatgg
ttaaaccagaattcacgtaaaccgactctaaacctagaaaatatctaaaccttggGATCCCAAGACCCTTCCTCTATATAAGGAAGTTC
ATTTCATTTGGAGAGG
```

TOC1

```
CAAAAACAAAAACAAAAACAAAAAgagaattttgttagcatgtcttcctcttctgggaccagtttgtgataaaacacatcctctcggaa
aagagtgtagagcacaacttcctctcgaatgtactaagaccactagactaacgtatagaagctctcaagtaaaatggctacgatccaaa
gagaatctgaaggtatgtgcaatgaggtcatgaaccatcatgatggtggtgataataacagattaacagcattgacaatttgaaaataa
tagtaatatgaacgcacaaatcatatttatttcttaaatagaaatgttttacaaaaacgattaatgtctaaattaattcaaggttctac
gaatGATAAGATAAGATAAGATAAaggtcagaatttgtatatgtagctaaattaaaataaataaataaacagatattttgtagaattgc
aaatatatgtgaataatcaaatataatagaacaagttggtcctcttcacatccttGATCCCAAGACCCTTCCTCTATATAAGGAAGTTC
ATTTCATTTGGAGAGG
```

CYP96A4

```
agaAATGTAATGTAATGTAATGTttttgattaaaggttaactacatttaacaattaacatatcggtatacactgaattaattgtgtaac
caccatttgtaaattatagtatcaacttttagatttatcatccgatatctttaatcacaaaaaataataaacgttatcacaggttactc
aaatcacatggAAATAAAAATAAAAATAAAAATAAttgcatattaccacaaagaagaaaagaacaaaaatcgaattgcattacacgtat
agtccaaaaatgcattatttgcttaaactgatattaagttagggataagcttgtctttctccacttaattaatatttttttgcatgttgt
gcaaactgatgtttcttaacaggttgtaagatcaatcaatacataataattataataataatcattcatcatgaattatttcatctaat
atttaattcatttggttgcctctcgtcagtttgtaatgaaaacatacaacaccaaGATCCCAAGACCCTTCCTCTATATAAGGAAGTTC
ATTTCATTTGGAGAGG
```

PRR7

```
ATATGATATGATATGATATGggataatcctaagtgtgtCAAAAACAAAAACAAAAACAAAAAtttaaactgagtactgtagatggaacg
atgtctctctgtgtcgtgggagggtggctcacagagctacttagtctaattatggtattgattaaaatacaaagatggagattattggg
taaggagaaaaaagtggaaggtgtggggatttatcttgaatcttttaatcttgcaattaggttaagtggcactgttggccaactaaaag
caaccccaagtatactcaataatgacactcataatctctttctccctctgtctattcctataatgacttaggaattacaaaattcttca
aggagataatgatactaacttgcttacaaagttataaaacatacacgacattgcaatataaaccctaccaaaatccctacttttccatc
atttcattcactcacacacgcataacatatttgtagcatccatacataaatacatGATCCCAAGACCCTTCCTCTATATAAGGAAGTTC
ATTTCATTTGGAGAGG
```

FIGURE 4.14: Synthetic promoters designed for representative genes of each of the four modules. The motifs are placed in tandem repeats in red, whereas the minimal promoter region is green.

the same classification was carried out one more time using slightly different designs. Coalitional game theory approach can be employed for that purpose as proposed in (Strumbelj and Kononenko, 2010). The method in a nutshell: 'Let $N = 1, 2, ..., n$ be a set representing n features, f a classifier, and $x = (x1, x2, ..., x_n) \in A$ an instance from the feature space. Let c be the chosen class label and let $fc(x)$ be the prediction component which corresponds to the class'(Strumbelj and Kononenko, 2010). The aim is to explain how the given motifs contribute to the prediction difference (between the classifiers prediction for this particular gene and the expected prediction if no motifs are selected for). The fundamental limitation of the present methods is that they do not take into account all the potential dependencies and interactions between feature values. Here this is overcome by defining interactions, by noting that each prediction difference is made up of 2^N contributions of interactions.

Essentially, the genes feature values form a coalition which causes a change in the classifier's prediction. One divides this change amongst the feature values in a way that is fair to their contributions across all possible sub-coalitions as discussed on other examples in (Strumbelj and Kononenko, 2010). If two features values have the identical influence on the prediction they are assigned contributions of equal size. If a feature has no influence on the prediction it is assigned a contribution of 0. For example, when looking at the particular contributions, both the magnitude and the sign is considered. If a contribution is greater than any other it is important for the entire model. Positive values can be translated as the particular feature increasing the model's output. A negative sign depicts that the particular feature contributes to decreasing the models output. Here caution has to be retained in terms of interpretations as clearly the absence of motifs is not as useful of information. One can delineate how much the model's output are altered when given the feature values for the instance, which features are responsible for this change, and the magnitude of an individual feature-values influence. That all leads to the prisoner's dilemma for the *cis* motifs detected beforehand (it is clearly one the other, both and none). It is interesting to phrase this cis motif problem in such a manner as then certain rules for understanding can be inferred particularly if a tool is build for that purpose and one can see the change in prediction given x number of combinations.

4.2.4 CIRCADIAN SYNTHETIC PROMOTERS AS PHASE HALLMARKS

The discussed analyses in this chapter led to synthetic promoter design for each of the four circadian classes. Motif editing and placement was carried out using PromoterCAD (Cox et al., 2013). Synthetic promoters can control the timing, location and amount of gene expression for any organism. PromoterCAD is a web tool intended for designing synthetic promoters with altered transcriptional regulation. Natural *cis*-regulatory motifs were placed into a synthetic promoter sequence. In theory the motifs can be copied into the corresponding position of the synthetic promoter sequence versus the genome sequence analysis can be used to predict functional locations of a motif. Both of the strategies can be combined with a strategy of stacking multiple copies of a motif, which aids to ensure that at least one copy is in a functional location. That exactly was carried out here. In a previous study carried out by (Ukai-Tadenuma et al., 2008), it was mentioned that 'We observed that reporter vectors with 1 or 2×UAS did not exhibit high-amplitude oscillations, while those with 3, 4, and 5×UAS produced high-amplitude oscillations in comparable phases. In this manuscript, we adopt 4×UAS because outputs with 4×UAS results in the relative amplitudes with the smallest variation'.

As it was shown previously, the minimal promoter region, 45 bp upstream of the transcriptional start site, is necessary but not sufficient for CaMV35S expression. This region includes crucial sequences for strongly regulated TATA-type promoters: the TATA box, a plant-specific CT-rich region called the Y-patch, and the initiator region surrounding the transcription start site. All these are maintained and the selected motifs from RF are included. This strategy implemented led to the design of such synthetic promoters ready to be tested, Figure 4.14 (and Appendix). These synthetic could now be synthesized, transformed into plants and measured for their temporal expression pattern (using a firefly luciferase reporter vector). In other words, a successful reconstruction of the natural circadian phases would imply that the network structure of transcriptional circuits determines the fundamental timing of the outputs from the circadian system.

4.2.5 DYCK PATH REPRESENTATION APPROACH ON INDIVIDUAL TREES AND FORESTS

Motivated by the outcomes of random forest it became interesting to experiment with tree data objects and PCA on those tree data objects directly, both ensembles resulting from random forest and single trees derived from the information on circadian features. The aim is to propose interaction models which describe the inner workings of a transcription system, on which further investigations may be based. The representation of trees in Euclidean spaces can be challenging yet carries a lot of potential in terms of analysis of the statistical properties of the tree data objects. Through the application of Dyck path representation (DPR) and tree pruning the tree data analysis allows for inference of interesting features, single and interacting, representative of a class (Shen et al., 2014). Recent publication of (Wang et al., 2012) proposed the framework enabling the development of adapted version of PCA for tree data objects (here specifically the typical Euclidean concepts pose a challenge). A major contribution was their devlopment of analog of PCA for tree data objects. This effectively led to representation of trees as curves. The bridge enabling that was DPR used to transform trees into Euclidean curves. The principle of Dyck path representation is presented in Figure 4.15. The obtained DPR curves are mean centered and then projected onto the main components. The class labels were permuted. The empirical p-value was used to assess differences between the classes (phases of expression). Such a method provides interesting alternative of interaction detection still generating non-intuitive hypotheses.

The trees were obtained from random forest. The objective of this method was to find the most occuring interactions characteristic for a studied group of interest. The results involved that every data tree object has been transformed into a vector and a corresponding matrix has been generated marking the position of nodes within each tree. That allowed for PCA on the new vectors. The code is deposited in the appendix and described in Chapter 2. Figure 4.16 represents a tree generated per class reconstructed

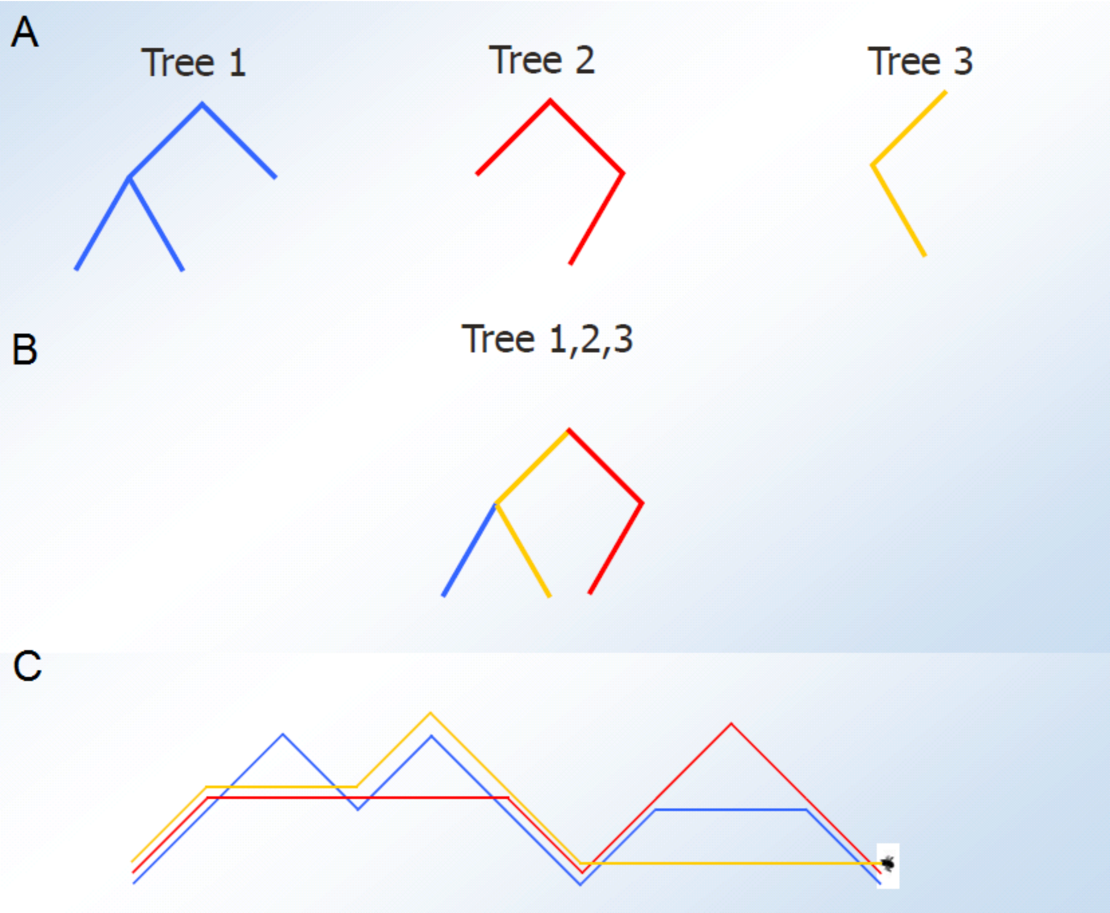


Figure 4.15: Dyck path representation principle. Panel A represents three example trees, Panel B represents combined trees whereas Panel C represents trees converted into curves based on a presence of similar node and lack of node. These curves (colour corresponds to tree) will be used as input for PCA.

for the motifs. It is representing first class versus the rest of genes classification of a frequency matrix so the splits are either more or less than x number of motifs. It has been pruned and explains at least 80 genes in a terminal node. The pruned support tree structure helps improve the efficiency of the statistical analysis as decreases the number of missing values on the tree. GraphViz software which was used for visualization of such a tree altogether allowing for exact inspection of interactions and presence of particular motifs. This specific example clearly depicts the logic principles such as having AATATC and TGGTGCAA at specified positions giving rise to the correct classification of a subgroup of genes. Furthermore, having the former, not having the latter and having TATCCA correctly classifies another subgroup. Given individual trees GO enrichment of these motifs can be and was analysed. One example of a multiclass classification is depicted on a tree where apart from the inherent properties, the GO categories of the groupings were considered. This shows the GO terms concurrently with elements resulting from a classification. Interestingly in one such tree present in SM, photosynthesis was grouped with the GGATAA and defence response has ACTTGCTG motif. This

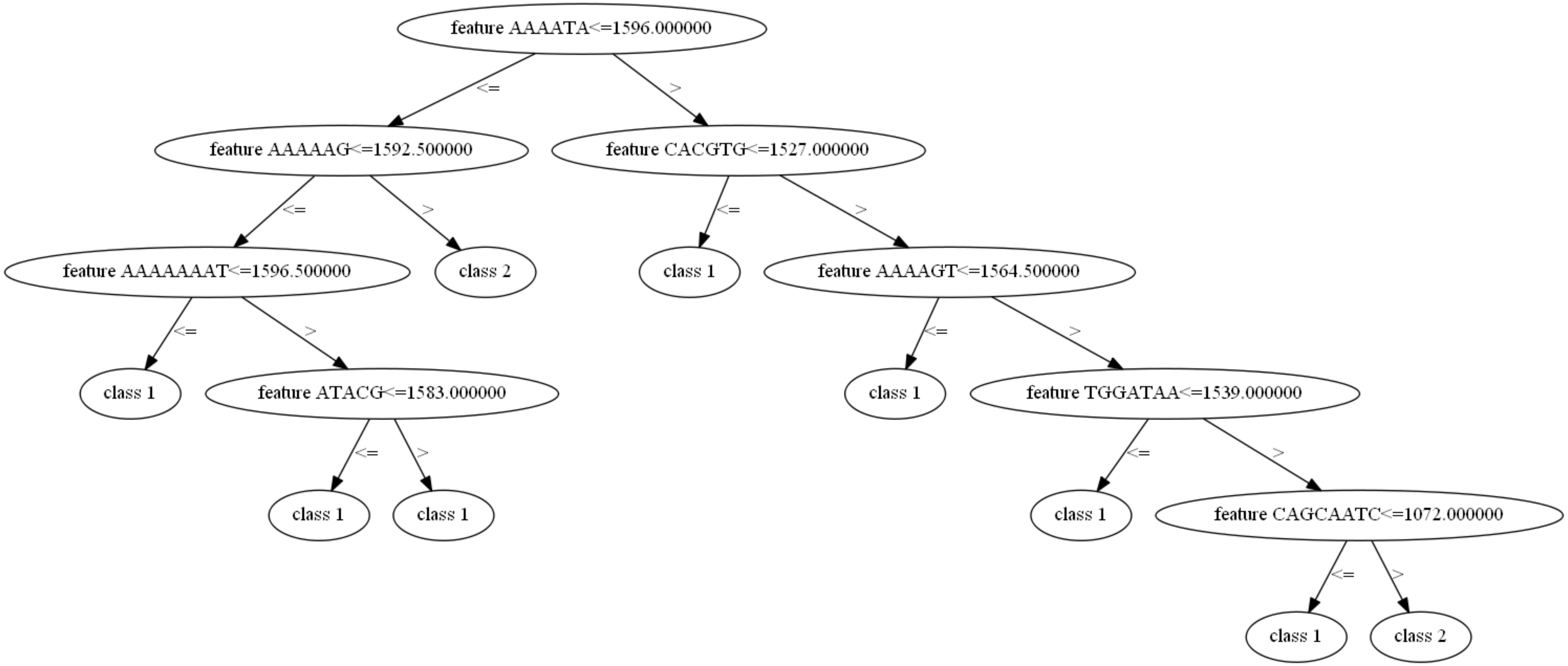

feature AAAATA<=1596.000000
<=
>
feature AAAAAG<=1592.500000
feature CACGTG<=1527.000000
<=
>
<=
>
feature AAAAAAAT<=1596.500000
class 2
class 1
feature AAAAGT<=1564.500000
<=
>
<=
>
class 1
feature ATACG<=1583.000000
class 1
feature TGGATAA<=1539.000000
<=
>
<=
>
class 1
class 1
class 1
feature CAGCAATC<=1072.000000
<=
>
class 1
class 2

FIGURE 4.16: Fragment of a decision tree showing how the presence (1) or absence (0) of promoter motifs and its frequency which was used to classify genes into phase 1 versus other phases. Here the most important split is AAAATA and ¡ ¿ represent either having or not having the given motif. The tree is pruned.

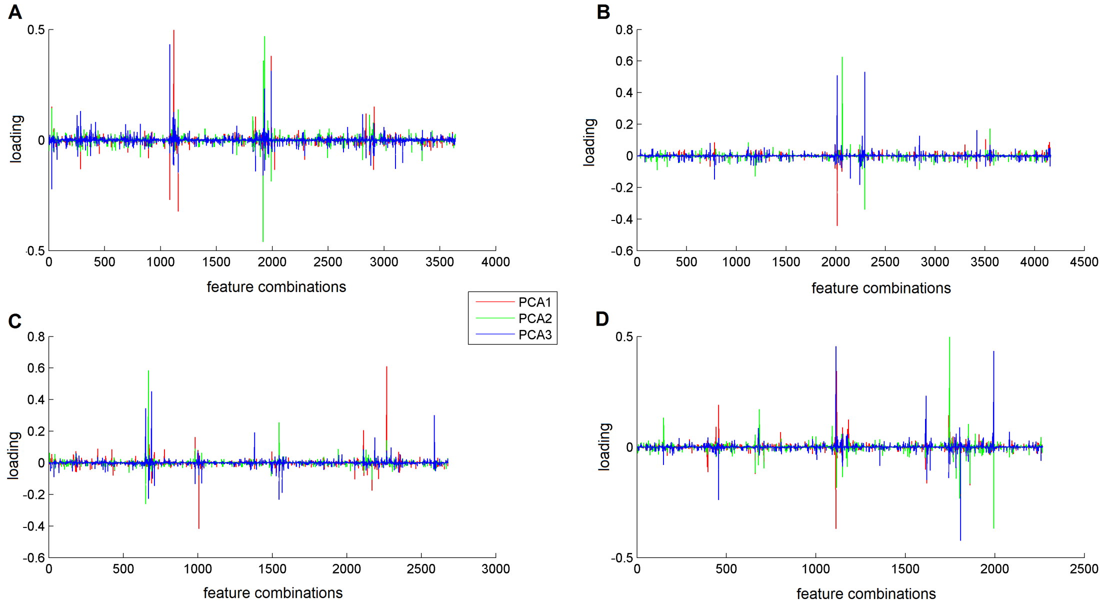
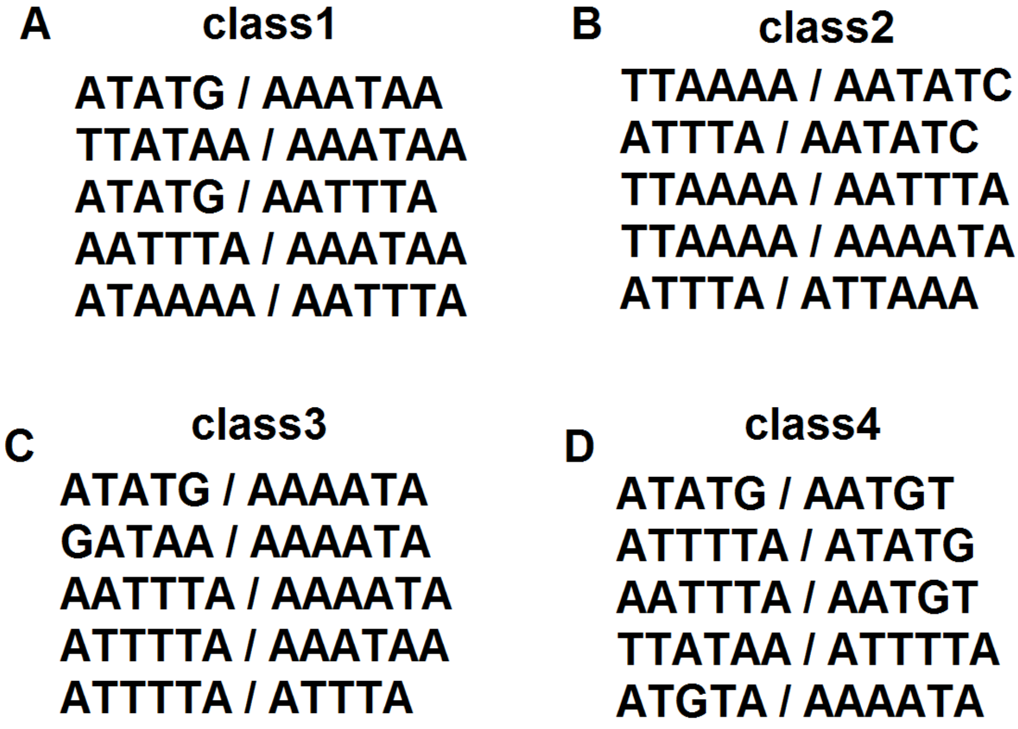


FIGURE 4.17: PCA on the population of the transformed trees given tree components. The x axis represent motif interaction and the y axis represents loadings. Further projection may be used to explore the loadings. Panel A B C D correspond to classes 1 2 3 4 respectively.

layers up another bit of information on the functionality of a specific class. Figure 4.17 represents PCA already on the generated curves. PCA was performed on 3638 x3638 pairwise non zero interactions. The curves centered and then projected onto the first component. Each eigenvector is a collection of features that carry the greatest variance. That shows the main modes of variation. The interesting phenomenon by studying the eigenvalues and eigenvectors is that there are many curves which are flat, and fewer which have much more structure. Altogether 1000 trees were generated, mtry=30, nodesize 1/8 of the size of the dataset. The zero values were removed across the 1000 trees per class. The resultant error rate per class is low with classes in order 1-4: 12, 13, 6.4 and 4.7 percent. That led to a series of motif pairs as highly important for each class. The applicability of the approach results from the direct ability to capture such interactions on a tree as opposed to the black box model. It would be interesting to add other features given alternative data as features to detect other aspects of the architecture. When the method was tested on 3 types of trees at the same frequency, a synthetic test set, the eigenvalue distribution reflected that in 2 PCA components explaining 100 percent of variation. The contribution of such method is that complex topologies can be directly analyzed and the vectors of interaction depicted in a unsupervised manner. On a presence absence matrix 1000bp matrix the error rate was 8, 8.7 14.1, 13.8 per class in order error rate. There was no initial preselection and the matrix consisted of 449 motifs. The motifs per all classes are presented in Figure 4.18. Whereas, here the method is used to detect strong interaction signals, it is equally applicable to any



A class1

ATATG / AAATAA
TTATAA / AAATAA
ATATG / AATTTA
AATTTA / AAATAA
ATAAAA / AATTTA

B class2

TTAAAA / AATATC
ATTTA / AATATC
TTAAAA / AATTTA
TTAAAA / AAAATA
ATTTA / ATTAAA

C class3

ATATG / AAAATA
GATAA / AAAATA
AATTTA / AAAATA
ATTTTA / AAATAA
ATTTTA / ATTTA

D class4

ATATG / AATGT
ATTTTA / ATATG
AATTTA / AATGT
TTATAA / ATTTTA
ATGTA / AAAATA

Figure 4.18: PCA output for four classes (established in Chapter 3). The output motifs consitute both previously annotated and yet unknown motifs).

tree oriented dataset including individual circadian genes and patient data integrating for example mutations, copy number variants and other features of importance to the study.

4.3 DISCUSSION

4.3.1 INTERACTIONS BETWEEN PROMOTER ELEMENTS THAT DETERMINE PHASE

Groupings of genes into four groups according to positive or negative correlation with these two components were used for a novel combinatorial search for promoter elements responsible for the phase of expression of circadian genes in Arabidopsis. We used the ensemble machine-learning approach, Random Forest. This decision tree method has the advantage of being able to consider the combinatorial contribution of multiple promoter *cis* elements to the overall expression pattern of a gene, an important consideration in the search for circadian *cis* elements in light of recent findings published elsewhere as to the nature of transcriptional control of the pattern of circadian gene expression in

Arabidopsis (Li et al., 2011a). The elements identified as most important in combination with others included a number of entirely novel elements but also some that are part of elements previously-identified as conferring phase information in isolation, such as the evening element.

We showed that, for each module, combinations of these promoter features could be identified as being able to provide sufficient information to allow genes to be assigned to the correct module with high precision and recall. Furthermore, a number of elements were common to more than one phase module group and these common element groupings changed in a progressive pattern from one phase group to the next so that removal of certain elements or addition of others could change the predicted phase of a gene in a consistent manner. It may be speculated that the final phase is determined by the additive effect of multiple transcription factors peaking at specific points throughout the circadian day. Such a system would allow a relatively small number of drivers to achieve a large distribution of phases, such as is seen for the plant ”circadiome”. In further support of the merit of our results, the previously-highlighted elements of interest occur in phase module groups consistent with the known function of the elements of which they form part. AATATC, part of the Evening Element occurred in the ”dusk” (class 2) phase group; GATAA, part of the I-box occurred in the ”noon” (class 3) and ”dusk” (class 2) phase groups; and CAAAA, part of the CDA-1 occurred in the ”midnight” (class 4) phase group.

Combining these transcriptomic findings with similar analyses based on microarray data from mutant genotypes would also be interesting and may shed further light into the architecture of the clock. The ultimate test of their veracity would, however, be to determine whether it will be possible to generate phases in cells using synthetic promoters containing combination of these elements. More precisely, to discriminate between promoters of genes belonging to different classes, that is modules relying on a set of features we were able to predict the given class with high confidence. That suggests the high information content of the features and the applicability in the design of synthetic promoters. As straightforward as it sounds, that called for several sets of features, careful module assignment and consideration of interacting features.

We additionally generated a novel classifier that relied on *in silico* occurrences of binding sites within their evolutionary conserved regions. We evaluated the models ability to accurately predict expression using the area under the receiver operating characteristic curve (AUC) and the CROC curve shown in Figure 4.11. To be specific, all of the models, we obtained AUC values higher than 0.8. Moreover, when we tested a model trained on a particular dataset on the promoters of genes expressed in another one, we obtained relatively high AUC values. This result suggests the existence of different subsets of promoters hallmarks, with characteristic sequence features. Modest performance is likely explained by lack of sequence features and/or relatively high heterogeneity of the promoters in the training set of the model. The models did not improve greatly by the use of frequency yet they did in terms of position. It is safe to state that the genes regulated by these promoters exhibit similar conservation trends, especially considering the extensive conservation across clades like for example seen in Brassica. In summary, our results indicate the existence of largely dissimilar sets of phases specific regulatory sequences. Finally, consistent with the hypothesis, the interacting features yielded even higher performance scores. The synthetic promoters that should be tried first are depicted in Figure 4.14 as they are representative of each of the four modules. Those particular motif dependencies could not be captured in any other method. A recent study presented a comprehensive comparison of RF to other methodologies including LDA, QDA, logistic regression, partial least square, KNN, neural network, SVM, and other classification methods using seven microarray gene expression data sets. RF was shown to have the best performance among all the tree derived methods (Lee et al., 2005). It has been applied with great success to predict protein DNA binding sites looking at amino acid residues. Yet it is not solely the choice of the algorithm here that helps yet understanding the properties like the behavior of attributes. It has been recently pointed out that interaction parameters in statistical modelling should be jointly interpreted with dominant effects for discovering biological interactions. RF and other treebased methods have a clear leverage here. Previously, Lunetta conducted one of the earliest simulation experiments to evaluate the power of RF to screen SNPs with interaction effects in GWAS and it proved that RF VIMP outperformed Fisher's exact test when risk SNPs were interacting (Lunetta et al., 2004). In terms of visualizations of single genes nomograms are useful in terms of showing the effects of *cis* motifs. Nomograms which are essentially risk calculators are being used for cancer prognosis, primarily because of their ability to reduce statistical predictive models into a single numerical estimate of the probability of an event. A nomograms are usually constructed from a set of clinical data which contain various attributes. It is of importance to select effective ones from available attributes, sometimes along with parameters accompanying the given attributes. Here nomograms are adapted to predict to defining characteristics of promoter regions. Up to now, the simultaneous effects on outcome were rarely examined, and this is one reason why some models, even with high sensitivity, lacked specificity (Iasonos et al., 2008). In the present work interactions were considered and the chapter laid foundation for *cis* nomograms that may be built on a grand scale using for example NETLOGOS and through incorporation of known *cis* elements from motif collections like JASPAR. The justification for the power of such approach has been provided by recapitulation of phases in the mammalian clock as shown by Ukai (Ukai-Tadenuma et al., 2008). With regards to other examples of cis interactions that were deeply analyzed, the extensive analyses of transcriptional regulation revealed that primarily E-boxes, D-boxes, and ROR elements control the rhythmic expression of genes. Such system provides examples for the interplay of two antagonistic regulators in the determination of the output phase, and illustrates the concept of synergy between activators and inhibitors. Some such interactors include ROR-REV-ERB, DBP -E4BP4 and BMAL1 -DEC1. Clearly the mRNAs and proteins of activators and inhibitors are often times expressed at opposite phases, in detail the maximum of activation coincides with the minimum of the inhibition. Yet similarly to the plant circadian system, for E-boxes, such a simplification does not apply. E-box regulation is governed by a combinatorial effect of at least 11 genes (BMAL1, BMAL2, CLOCK, NPAS2, CRY1, CRY2, PER1, PER2, PER3, DEC1, DEC2) (Korenčič et al., 2012). That is further complicated by numerous post-translational modifications as elegantly shown by E-box regulation. It can be considered a hub of the circadian gene regulatory network similarly to the EE core sequence. The combination of the RREs with the D-boxes and the E-boxes may lead to phase shifts in both directions. The underlying design principles include the overcritical delays which are crucial to obtain oscillations, and synergies of activators and inhibitors that enhance amplitudes (Korenčič et al., 2012). Such concept in plant biology carries even greater potential as the clock exerts pervasive control affecting majority of physiological outputs.

In terms of purely biological outputs the plant *cis* elemtome is largely unexplored. There were indeed several such investigations yet focusing on the upstream promoter regions of individual genes, not the entire *cis* regulatory networks. Such studies like for example by (Michael and McClung, 2002), not to mention the foundation laid down by (Harmer, 2000) were of great importance to the establishment of core databases. Nevertheless, aiming at the system wide explanation tells a lot about the *cis* motifs which reflect back on the genes. For example the EEs at different positions and frequencies in specific combinations are certainly a hallmark of rhythmicity.

4.3.2 MULTIPLE FEATURES TO PREDICT BINDING SITES

The problem at hand is that there are several tools that focus on motif prediction yet with their advantages they are not adapted to solve the problem at hand. Here through series of tests the features resulting in the greatest information gain were preselected and used to delineate the key sites of circadian regulation. The aim is to learn a function that maps a set of features measured for a location in the plant genome to a score indicative of transcription factor binding that location.

$$P(b|f_b) = \frac{1}{C} \times \sum_{c=1}^{C} P_c(b|f_b) \tag{4.2}$$

We divided up the genome into a list of genomic regions pertaining to the promoter. Each region has a corresponding binary label 1/0. Each region is described by a set of features. The aim of the learning is to find a function f (called a model) to define a region is a fully functioning motif from some observed features alone. That is similar in principle to CONSURF and the tools used for example by the UC Genome Browser nevertheless adapted for plant genomes (Ashkenazy et al., 2010). To find a suitable f, one has to decide on a mathematical form of f, identify known positive and negative examples that can help estimate the parameters of f; and actually estimate the parameters of f, in a way that it likely predicts the labels of regions accurately, even for regions for which the corresponding labels are unknown. Random forest was used for task 1. Procedure-wise, over-fitting is detected by building a model based on a subset of the examples (the training set), and evaluating its accuracy based on another subset not involved in training (the testing set). An over-fitted model would have good training accuracy but poor testing accuracy. That was accounted for by cross validation. The idea that label of each genomic region can be determined by its own set of features alone is not entirely solid. Features are likely dependent. Nevertheless a series of features were used and a sliding window approach to define the regions of greatest likelihood of binding. The code is available in the Appendix whereas the results were used for classifications. The code can certainly be used with weights attributes with respect to the problem at hand across all plant species. To uncover the putative binding sites we:

- Divided the dataset of motifs into training and test (2/3, 1/3) with 2 classes
  - ”Real” binding site:
    - Jaspar Core Plantae
    - Literature curated
  - ”Negative” dataset
    - All the windows that are not similar to real binding sites
    - Windows of 20bp in size
- Selected for a series of features looking at information gain:
  - Conservation
  - Average estimated melting temperature
  - Percentage of GC content, 50bp each direction

  - If window/motif is within 3' / 5' region
  - Nearest ATG in bp

- For the feature 'Conservation', all 3171 "circadian" genes, 1200bp upstream sequence extracted in FASTA format. Orthologs in other species according to PLAZA 2.0
  - *Brachypodium*
  - *Glycine max*
  - *Lotus*
  - *Arabidopsis thaliana*
  - *Manihot esculenta*
  - *Medicago truncatula*
  - *Oryza sativa*
  - *Carica papaya*
  - *Populus trichocarpa*
  - *Ricinus communis*
  - *Sorghum bicolor*
  - *Zea mays*
    - * Perform multiple alignment using ClustalW. The rate of evolution at each bp is calculated using empirical Bayesian paradigm. Each window will have an average conservation score

- For the feature 'Melting Temperature'
  - $Tm = \frac{64.9+41\times(yG+zC-16.4)}{wA+xT+yG+zC}$
  - wA - number of As
  - xT - number of Ts
  - yG - number of Gs
  - zC - number of Cs

- For the feature GC content each direction 50bp
  - Window size expanded to $120(50 + 50 + 20)$
  - $GC_{50bp} = \frac{G+C}{A+T+C+G}$

- Feature: Nearest ATG
  - Distance in bp from nearest translation start site

Considering the avalanche of data, the above can be supplemented by series of features other than the ones pre-specific, additionally user should be able to preselect features considering information gain. Even though the tool is not running it was useful to see depletions on the promoter sequence given such a classifier.

4.3.3 CONCLUSION

These core features like the motifs, their frequency and position together with network-scale characteristics of circadian genes (such as degree, connectivity, centrality, edge length, closeness) led to creation of several feature matrices that were used as training datasets to conduct a series of machine learning and network scale analyses. That set of results combined with the methods in the upcoming chapters should give insights on improved design of circadian systems, prediction of key circadian genes, prediction of key modules their preservation across species, crosstalk between these circadian modules and output pathways and the potential candidates for alteration and maybe even improvement (better adaptation) of the circadian systems in crops. Clearly computational analyses and understanding of how the biological systems process information can develop improved models and algorithms and provide a more thorough explanation of how and why the system operates as it does. Ideally, the aim is to visualize the contribution of features looking at the change of predictions in interactive manner. That has been started together with nodetrix that allows user to interfere in networks. Tools like NETLOGO and the RF site prediction can aid biologists trying to solve other conundrums and hence the mission is the integrate them with iPLANT collaborative in the near future (Goff et al., 2011). Up to this day the models focused on individual elements like for example (Michael et al., 2008), whereas no intergrative *cis* regulatory model was presented in plants. Analogy can be made between the interchange of a network and intentional breeding. It is improving on nature based on observation, here focusing on the parameters of the entire system.

Chapter 5

THE STATIC AND DYNAMIC PROPERTIES OF THE CIRCADIAN NETWORK

5.1 INTRODUCTION

The focus of this chapter is on network analysis of the circadian clock, particularly differential interactions in both regular circadian Arabidopsis data and in a mutant dataset. In the recent years as NetworkBLAST and many other tools and methods to compare networks across species became available, more insights have been made into properties of the networks that are similar across aspecies. In the current chapter, the attention is on networks of Arabidopsis genes, their remodelling of as a result of environmental or circadian changes and due to the particular mutant states. Networks are quantitatively analysed involving varying genotypes, changing conditions and timepoints.

As network mapping methods are becoming more robust, high throughput and quantitative, major advancements are allowing exploration of the true interactome space. Regulatory networks clearly reconfigure dynamically as a function of specific context in which they operate. Analysis of static networks did establish the framework for understanding the plant circadian clock yet clearly to delineate and capture the circadian architecture, it will be essential to decipher the dynamic interactions. It will be interesting to elucidate whether different interactions happened at different times of the day, if they are recurrent at particular phases and how so. Previously, not much has been done in the particular field apart for few attempts that were though of great significance, like the study described by (Bandyopadhyay et al., 2010)

Network science can directly help understand the plant circadian clock structure, development and weaknesses (Alon, 2003, Barabási, 2013, Barabási and Oltvai, 2004, Boccaletti et al., 2006, Ghoshal and Barabási, 2011, Liu et al., 2011, Newman, 2003).The structure shows the properly functioning families of TF affecting expression patterns, their development may show changes in the network across time like does growth of a plant and differences across tissues, altogether giving insights into the evolution of such networks on a grand scale. The weaknesses of the circadian network may be translated into points where genome editing might be of use, the regions of the network where knocking out a hub would confer desired changes to the output of a TF for example. It was shown in yeast PPI network that the hubs are most likely to be lethal and there are many other examples (cite Jeung 2001). The organization of genes within the genome may reflect a mechanism by which proximity facilitates regulation. It has even been proposed that the coregulated genes involved in a common function contain a shared linear position within the genome cite Kosak 2004. The circadian clock as many other real world networks takes a shape of a scale free network. The degree distribution follows a power law and a proof is presented later in this chapter. Such a circadian network is held together by a small number of highly connected hubs (genes like TFs that connect to many other genes), which is its highly important organizing principle. Another important organizing property of this circadian network is the "small world" property, which translates into genes being able to reach one another by a small number of steps. In such a network the distance between two randomly selected genes grows proportionally to the logarithm of the number of nodes in the network. Evolutionarily, such networks are likely more robust to perturbations hence seen in many other biological networks other than circadian. It likely expands and grows through preferential attachment which is common to both natural and synthetic networks. In such model the proability is greater that a new connection will be made with a hub gene. In other words, the hubs in such a network emerge and become larger over time explaining the scale free property. In addition, the circadian network is likely going to be robust to decay (nevertheless, collapse of the hubs could still have disastrous effects). Interestingly to the circadian field, it has been shown that preferential attachment increases the number of coupled feedback loops cite Kwon 2007. There is a correlation between the number of feedback loops and network robustness shown based on simulation in this study. Transcription is not solely a process that turns on a specific gene, but a process that affects within the genome an entire network of genes. Furthermore robustness depends upon the potential existence of paralogues as they can mediate the effects once alternatives are not functional. For example, in the case of a TF, compensation by a paralogue would mean that loss of a component does not affect the expression of target genes. This situation would be far different when knocking out a single acting TF which had no paralogue. Evolutionarily, it is possible that a backup system naturally emerged as a result of the process of duplication and divergence. The challenge, from the design standpoint when building a synthetic network, is to name the nodes that should be replicated to decrease detrimental effects of nodes and edge failures. The fundamental set of laws and mechanisms that governs the circadian network is critical to the understanding and modification that will soon be attainable. Quintessentially the scale-free networks reflect a kind of self-organization that is governed by its inherent substructure. Such characteristics make network theory a realistic model with which to approach the genomic organization of the transcriptional regulation. This chapter aimed to identify the core properties of the *Arabidopsis* circadian clock network represented in the form of coexpression matrix and then to identify the changes that are present in that matrix through the course of the day and night. These changes are identified by a method similar to differential interaction mapping (Bandyopadhyay et al., 2010, Bean and Ideker, 2012). The hypothesis is that such a method will allow one to infer the disturbed genes and modules and, hence, the GO disturbed processes. Modules are the building blocks of cells (Fortunato, 2010, Hartwell et al., 1999, Langfelder et al., 2011, Lorenz et al., 2011). At the molecular level, modules can be defined as groups of genes, gene products or metabolites that are functionally coordinated, physically interacting and at times co-regulated. That will reflect upon which genes associated with given phases are exerting their roles in the circadian processes. Having done this, microarray data generated for a circadian mutant was also analysed, essentially translating genotypic combinations to phenotypic manifestations.

Datasets used are the same circadian Arabidopsis microarrays as mentioned in the chapter 3 and 4 (described in methods Chapter 2). The mutant dataset has been generated in the lab of Dr Paul Devlin as the focus of the lab was on *far1 fhy3* single and double mutants. Seedlings were grown in 12-hour light 12-hour dark rhythm at 21 degC for 1 week, transferred to continuous red light, grown on MS medium with sucrose, light intensity of 150 μmol m^{-2} sec^{-1}, samples taken at 8 hour intervals, 4 time points (24, 32, 40, and 48). The transcription factors FAR-RED ELONGATED HYPOCOTYL 3 (*FHY3*) and its homolog FAR-RED IMPAIRED RESPONSE 1 (*FAR1*) encode two proteins related to Mutator-like transposases. They act together to modulate *phyA* signalling by directly activating the transcription of *FHY1* and *FHL*. The expression of *FHY3* and *FAR1* is negatively regulated by *phyA* signaling (Lin et al., 2007). They bind FBS defined as CACGCGC. *FHY3* also co-regulates a number of common target genes with PHYTOCHROME INTERACTING FACTOR 3-LIKE 5 (*PIL5*) and ELONGATED HYPOCOTYL 5 (*HY5*) (Wang et al., 2011). *FHY3* is involved in regulating multiple levels of plant development. Foremost, the *FHY3* and *FAR1* are essential for circadian gene expression and flowering-time control (Li et al., 2011a). Here a qRT PCR analyses revealed that cyclic expression of *CCA1*, *LHY* and *ELF4* was decreased in *fhy3* and *far1* mutants in continuous white-light conditions.

5.1.1 IDENTIFICATION OF MODULES

Modularity is an important property in biology because it helps a system 'save its work' while allowing further evolution (Hartwell et al., 1999, Kitano, 2004). Modules can be defined by a group of nodes that are more strongly intraconnected than interconnected, as a number of edges inside a subgraph which exceeds the expected number of internal edges that the same subgraph would have in a null model; and as a set of coexpressed/co-regulated genes; and as a set of conserved sequences/structures (Fortunato, 2010) Modularity contributes positively to fitness by several indirect means. These are the instant benefits of modular organization in such a circadian system. Modular systems are more robust because the effect of perturbations can be contained within a module (Kitano, 2004). Modularity also enhances evolvability because it allows different parts to be optimized separately without impairing the functioning of other parts. Furthermore, a rewiring of modules can be achieved quickly in response to environmental perturbation. Finally, modularity makes the exchange of genetic information much easier. It is present in genetic, metabolic and protein protein interaction networks. Systems which are directly involved in interactions with the environment would evolve to be more modular than systems which have no external interactions (Singh et al. in an evolutionary study of three bacterial stress response). Modularity in gene regulatory networks decreases the complexity of the circuitry essential for complex responses to external stimuli. Transcriptional regulation factors are often acquired through horizontal gene transfer and it has been observed that regulatory circuits evolve faster than the genes they regulate (Boccaletti et al., 2006). Han et al. studied the modular structure of protein networks in more detail by considering their temporal changes. They found two types of hubs which they named "party hubs" and "date hubs" (Chang et al., 2013, Han et al., 2004). While party hubs interact with most of their partners at the same time, date hubs bind their partners at different times or places. They investigated roles of hubs played in the protein network. They observed that party hubs function mainly inside of modules while date hubs act as global connectors between modules. The theory of spontaneous emergence of modularity states that systems become modular under three conditions These include changing environments, information exchange and slow evolution. These conditions appear to be met in much of biological evolution. Mathematically, modularity measures the compartmentalization of biological organization. In the form of a linear expansion, the theory of spontaneous emergence of modularity would be stated as the rate of change of modularity is proportional to the environmental change. Whether such a building block is the 'final product of algorithm without a priori definition' versus 'a posteriori definition of a community' several methods were used to detect the modules and these were further assessed and comapared. Initially modules were detected on the static networks. These are made up of coexpression networks where the nodes are the genes and the edges show similar expression patterns according to a chosen similarity metric. These detected interactions may be representing functional relationships, physical interactions (as our and other overlay experiments have shown) and moreover logical relationships. The obtained coexpression networks do exemplify scale free topology and exemplify community structure similar to other microarray derived networks in biology (Barabási and Oltvai, 2004). Because it is likely that systems like the clock would benefit from modular organization, the initial aims in this chapter constituted the determination of existence of such modules. If modularity is the order in such a potential system, the aim is to capture it. Modularity can reduce the task of searching the entire space of possibilities into a polynomial problem of searching in the subspace of modular solutions. Modularity can be viewed as a reduction of pleiotropic effects. The aim was to detect, assess significance, compare partitions, analyze them using GO enrichment of nodes and edges and reconstruct them using a contribution matrix. Modular structure is pervasive in many complex networks of interactions and sheds light on the relation between the structure and function of complex systems. Furthermore every module can be seen as a collection of contributions of links from any node in the network. The challenge was to disentangle these contributions, to understand how the modular structure is built. The main problem is that the analysis of a certain partition into modules involves, in principle, as many datapoints as the number of modules times number of nodes. To confront this challenge, here we first define the contribution matrix based on the same dataset as a map of nodes versus modules to describe the partitions of relevance, a method to isoloate modules.

In this analysis a novel statistical approach for the differential analysis of molecular associations within the circadian system was introduced. The method relies on Pearson correlation of transcriptomic data quantifying the underlying interdependencies between genes. Using these measures of association we construct the differential networks and explore their features based on the networks topological properties. Large-scale genetic interaction networks have proved extremely powerful for mapping the pathways that regulate essential cell functions (Ideker and Krogan, 2012). In this study, we have shown that differential genetic networks are comparable in size to static networks as results have shown, yet access a very different set of interactions governing dynamic responses like in the time series circadian experiments. Given that most gene functions arise in response to changing conditions, the differential network revealed here offers a glimpse into a much larger universe of genetic interactions that are condition, cell type and potentially tissue and mutant specific. It is likely that the genetic interactions between these identified modules are reprogrammed in response to perturbation.

5.2 RESULTS

The results section of this chapter encompasses several stages of network analyses that will tie together and that flow in a meaningful order. The results section commences with static properties analysis 5.2.1 which is followed by community detection methods 5.2.2 section and module detection section 5.2.3, module enrichment 5.2.4 and overlays 5.2.6. These first steps feed into the later sections of the chapter that is the differential network analysis 5.2.6, time specific 5.2.7 and mutant specific 5.2.8. Whereas each stage is interesting in itself the flow is intentional as along the way comparisons between networks are being made.

5.2.1 STATIC PROPERTIES

To quantitatively describe and explore features of the circadian networks, we examined several topological properties. Of foremost interest was the degree, k_i, which is the network measure that indicates the number of connections that a node i has with the other nodes in the network. The degree distribution $P(k)$ represents the probability that a node i has k links. $P(k)$ can be calculated by computing the total number of nodes with degree $k = 1, 2, \cdots$ and dividing by the total number of nodes N. The betweenness centrality, b_i, is the measure that indicates how central a node v is in the network. The clustering coefficient, c_i, indicates to what extent the neighbors of a selected node i are connected to each other. Initially the core characteristics of such a circadian network were measured across static experimental design and assessed. One usually measures the connectivity of a node by its degree. It is useful to demonstrate the centrality of a node in a network. The retainment of links is subject to Pearson correlation being above 0.8. The datasets used were all the Arabidopsis circadian experiments discussed in Chapter 3. The results are depicted by Figure 5.1a through 1d and described below. This figure shows the global circadian properties. The degree distribution of the network provides the probability that a randomly chosen node has degree k. The distance distribution provides the probability that two randomly chosen nodes have a distance d between them (shortest path). Subgraph C portrays the degree distribution plotted on a log-log plot. Subgraph D shows the dependence of the average clustering coefficient on the nodes degree, k. This function is measured by averaging over the local clustering coefficient of all nodes with the same degree k. The results are comparable to a non random biological network of that size derived from gene expression data. They rely on the core circadian dataset and lay the foundation for further experiments. Theses results are particularly important in the light of the forthcoming differential interaction analysis. The scale free topology is demonstrated using the present gene expression data. These portray the degree distribution, the distance distribution and the dependence of the average clustering coefficient on the node's degree respectively. The static coexpression network served as starting point for the identification of modules using a series of methods, module comparison across datasets and species using several metrics, module enrichment looking at both nodes and edges and using overlay confirmation, specifically using BIOGRID interactions and PPI overlay from one particular experiment type (Stark et al., 2011). Throughout the text multiple comparisons were made between the static networks and differential networks as solely these allow one to capture the dynamics of a system.

5.2.2 COMMUNITY DETECTION METHODS

Given the coexpression networks, the objective was to determine all the modules within gene coexpression which will be compared. For that purpose several community detection methods were used. The dominant methods employed to detect modules were the eigenanalysis method described in chapter 3, the TOM (Topological Overlap Measure) methodology as described initially presented by Ravasz et al., and ClusterONE (Nepusz et al., 2012, Ravasz et al., 2002, Zhang et al., 2005). It is a standard to use the Pearson correlation coefficient as the major coexpression measure and that was of course used initially (Eisen et al., 1998, Holter et al., 2001). That brings one to focus on the proxy measures that essentially are the dot product with a specific eigengene, and the TOM score. The approach discussed by Ravasz relies upon the measurement of how close pairs of nodes are in a network (Ravasz et al., 2002). Essentially these first order interactions form the groundwork for measuring pair wise similarity. Initially the adjacency matrix $A = [a_{ij}]$ which encodes whether/how a pair of nodes is connected was constructed. The Pearson correlation matrix was dichotomized for that purpose as the operations were performed on an unweighted network. The standard threshold used was the 0.7 cutoff. The threshold was formerly chosen using the scale free criterion which suggests the presence of hub genes and robustness to random perturbations (Barabási and Oltvai, 2004). As thresholding might lead to information depletion, concurrently a weighted network was transformed using a power adjacency function from correlation to adjacency. Modules being here densely interconnected, genes were delineated using dissimilarity reliant upon topological overlap. This was performed in conjunction with the average linkage hierarchical clustering. The TOM measures the set of nearest neighbors of nodes and is defined by

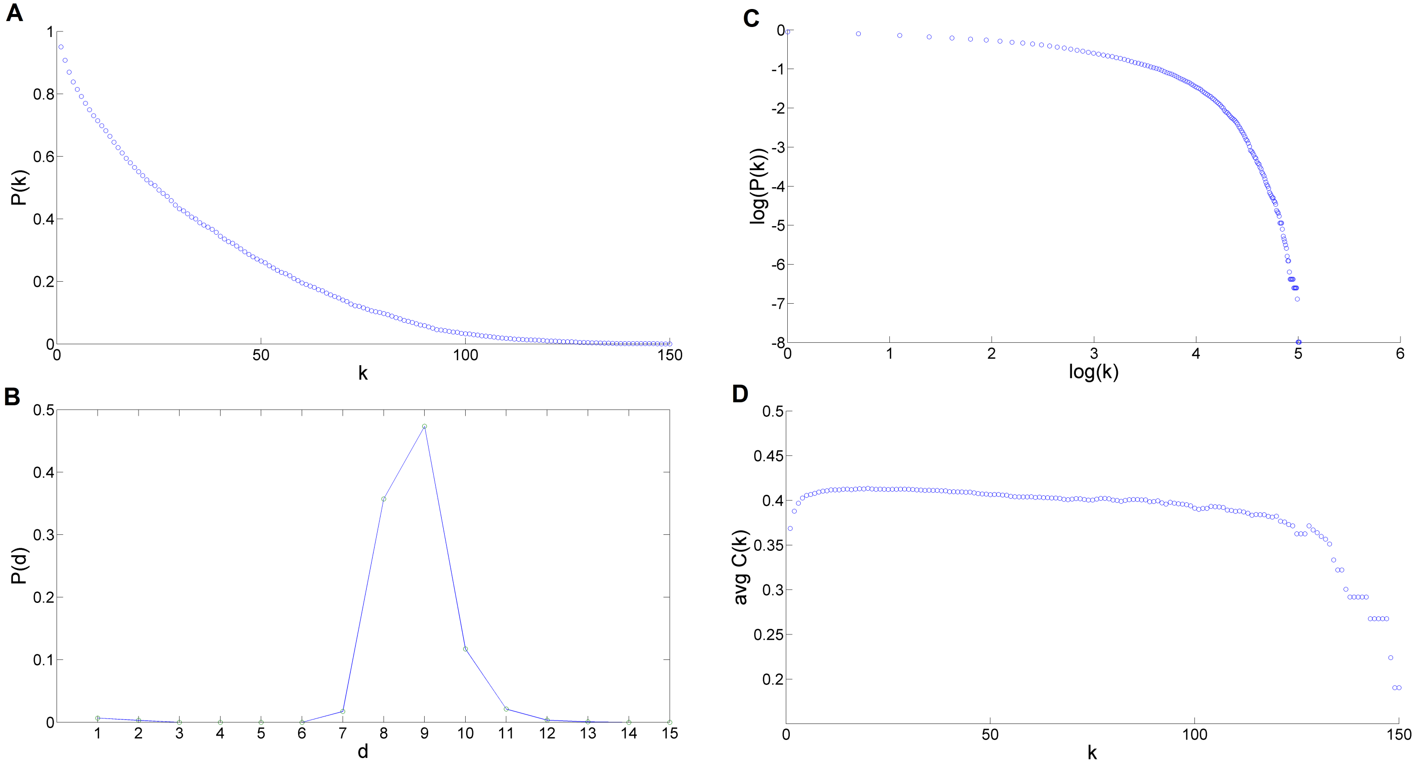


FIGURE 5.1: Global circadian properties. A. The degree distribution, p_k, of the network, providing the probability that a randomly chosen node has degree k. B. The distance distribution, p_d, providing the probability that two randomly chosen nodes have a distance d between them (shortest path). C. Shows the degree distribution on a log-log plot. D. The dependence of the average clustering coefficient on the nodes degree, k. The $C(k)$ function is measured by averaging over the local clustering coefficient of all nodes with the same degree k. The correlation cutoff is 0.8.

$$TOM_{ij} = \frac{\sum_u a_{iu}a_{uj} + a_{ij}}{min(k_i, k_j) + 1 - a_{ij}} \tag{5.1}$$

$$DistTOM_{ij} = 1 - TOM_{ij} \tag{5.2}$$

This helps alleviate the multiple testing problem (ambiguity) of finding genes significantly correlated with phenotypes (Langfelder et al., 2011). Multiple testing correction adjusts the individual p-values per gene to keep the overall error rate to less than or equal to the prespecified p-cutoff value. The TOM method as compared to GTOM defined as the Generalized Topological Overlap Measure, which favours larger modules, was used and that was particularly important as the major difference between TOM and GTOM is the incorporation of higher order interactions in TOM.

The obtained modules using the circadian Arabidopsis gene coexpression dataset were compared and inspected for enrichment of particular terms. Figure 5.2 portrays one such dendogram obtained for a WT set with all the genes. That as expected gave rise to many clusters and hence the same experiment was repeated on a smaller group of genes as the next figure describes. One can only detect changes across conditions knowing the first hand modules in the WT set. Figure 5.3 portrays the obtained groupings when done on the preselected group of circadian genes. As the particular module network may encode a pathway, these special types of networks have great practical importance. We used the functional gene annotation tools from the Database for Annotation, Visualization and Integrated Discovery (DAVID) to test for both GO enrichment and KEGG pathways. Interestingly, these findings are in agreement (section 5.2.3) with major clusters detected using eigenanalysis yet create a higher number of partitions. One could map the 4 previously described modules. The colours on Figure 5.3 are simply demarcating the boundaries between the groupings.

Subsequently the last method used for that purpose of module detection was ClusterONE. It is a graph clustering algorithm that is able to handle weighted graphs. It readily generates overlapping clusters. In theory, a subgraph representing for example a protein complex should satisfy two simple structural properties: it should contain many reliable interactions between its subunits, and it should be well-separated from the rest of the network (Nepusz et al., 2012). These two properties were taken care of in a quality measure that was called cohesiveness. The algorithm essentially detects complexes from weighted networks, using cohesiveness to guide the search. In a nutshell the algorithm grows groups with high cohesiveness from selected seed proteins, starting with highest degree first seed and grows a cohesive group from it using a greedy procedure (Nepusz et al., 2012). In the next step, the algorithm selects the next seed by considering all the proteins that have not been included in any of the protein complexes found until

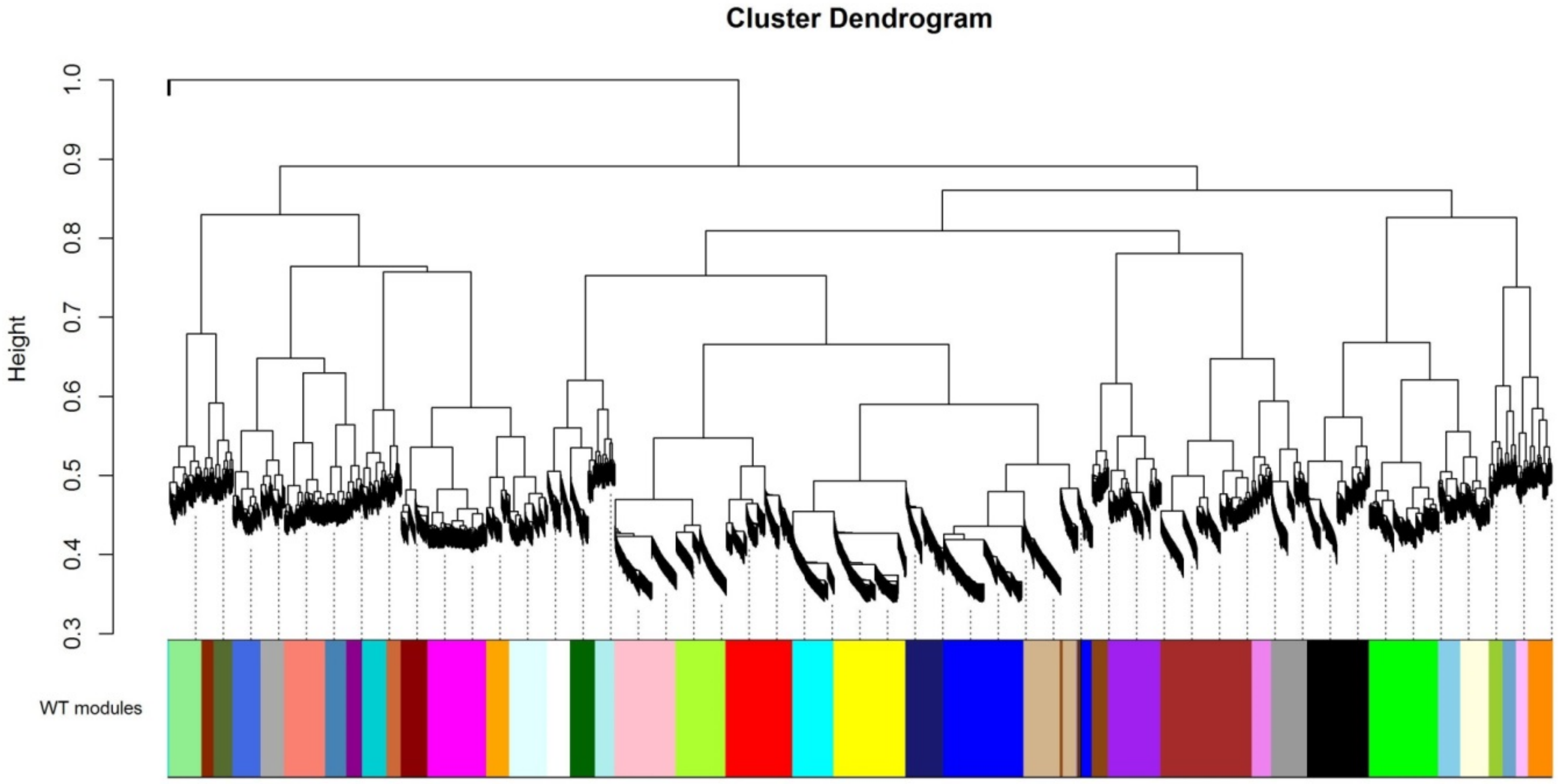


FIGURE 5.2: Clustering dendrogram of genes, with dissimilarity based on topological overlap (WT, 33 modules), that is one example of module detection. Module dendograms show clusters of modules with high coexpression.

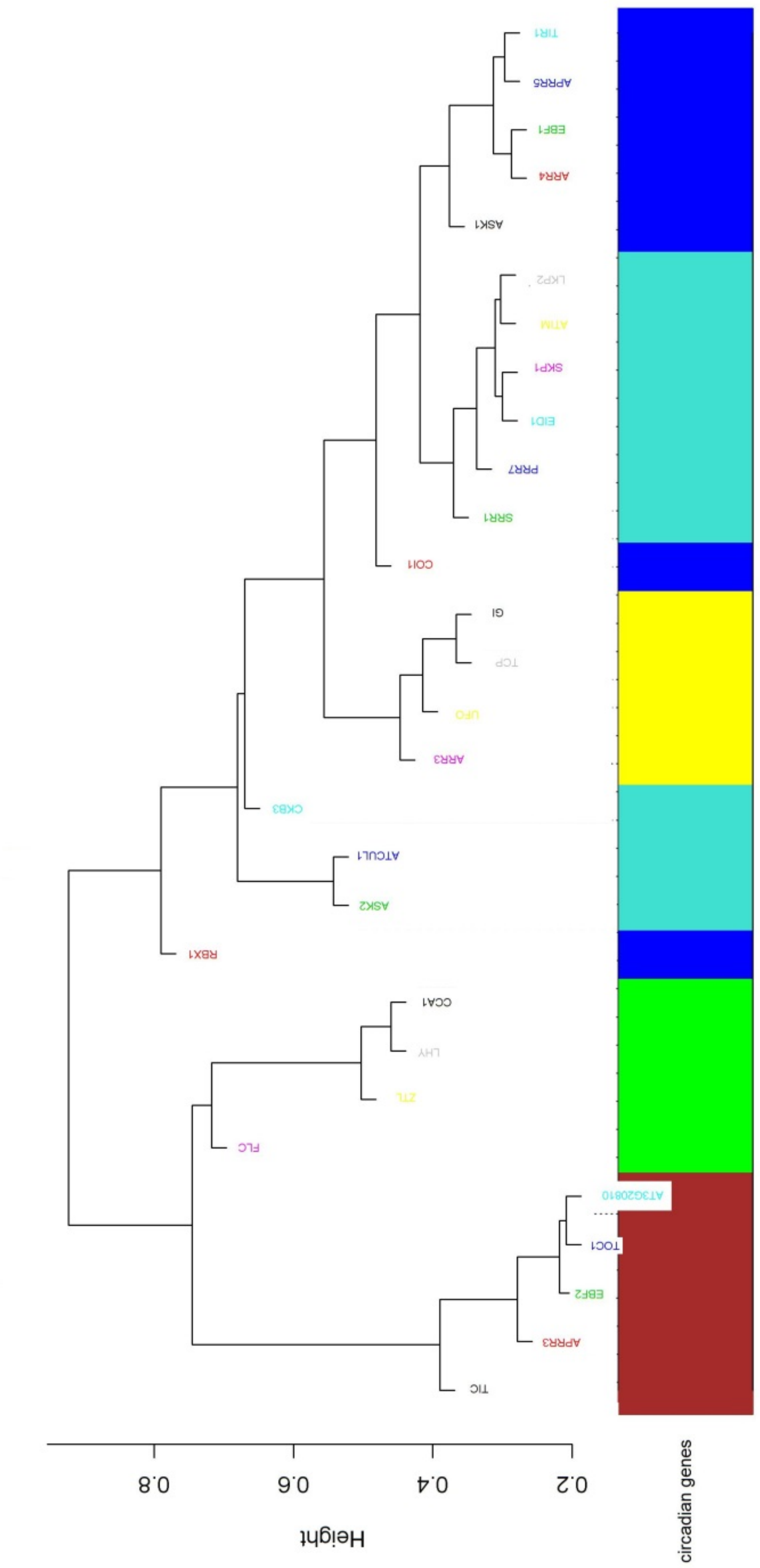


FIGURE 5.3: The dendrogram results from average linkage hierarchical clustering. Clustering dendrogram of genes, with dissimilarity based on topological overlap with GO Circadian clock genes forming the input nodes. The colour-band below the dendrogram denotes the modules, which are defined as branches in the dendrogram. Altogether eight clusters were detected with *CCA1*, *TOC1*, *GI*, *APRR5*, *APRR7* and *COL1* representing individual clusters.

now and selects one with the highest degree. The entire procedure terminates when there are no proteins remaining to consider. The ClusterONE was implemented on the same circadian set in the form of adjacency matrix (derived from coexpression) and the results are depicted in Figure 5.4. It portrays the obtained clusters of which many overlap (altogether two overlapping with the eigenanalysis modules and two subdivisions). The ClusterONE overlaps were merged and that gave rise to six modules portrayed on Figure 5.5 and Table 5.1. Figure 5.5 portrays inter and intra module interactions. It is clear that the greater number of interactions will happen within the modules and that is the situation yet it is even more interesting that the sole other interactions were

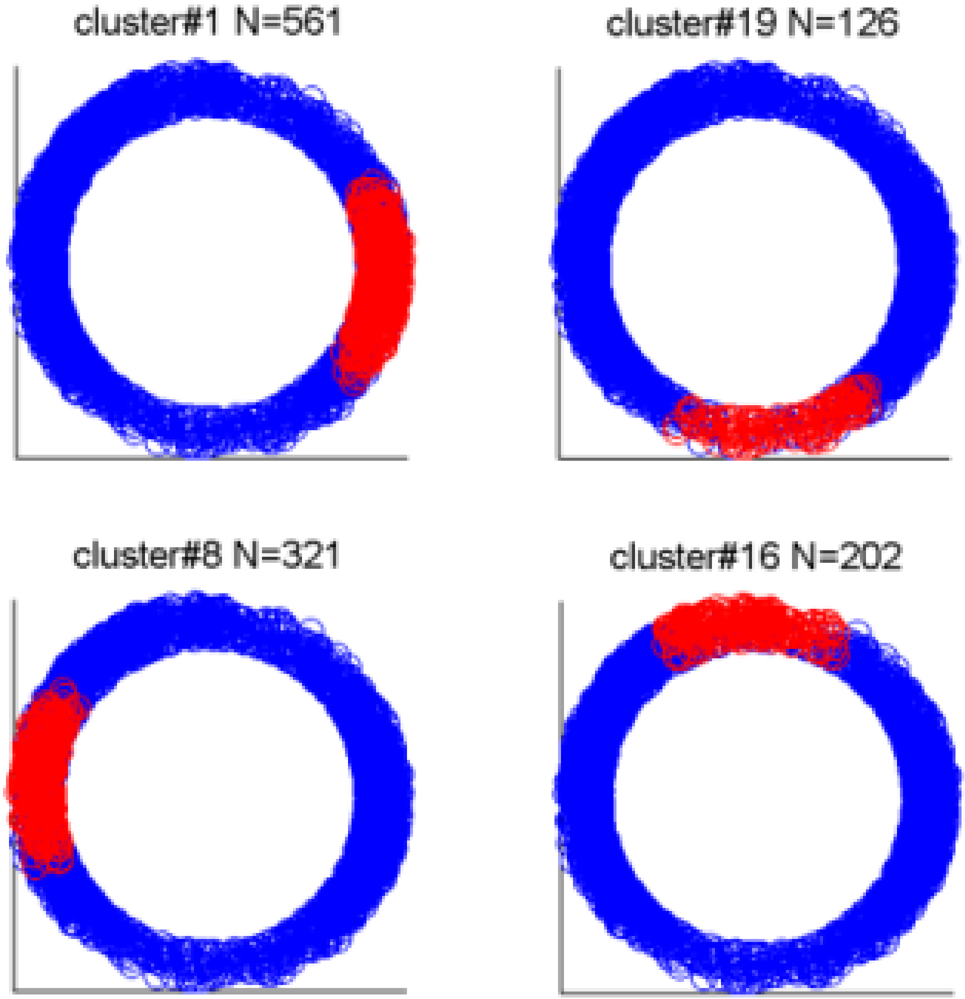


Figure 5.4: ClusterONE raw output shows the four potential modules, which served as input for further analysis. The algorithm allows for overlap.

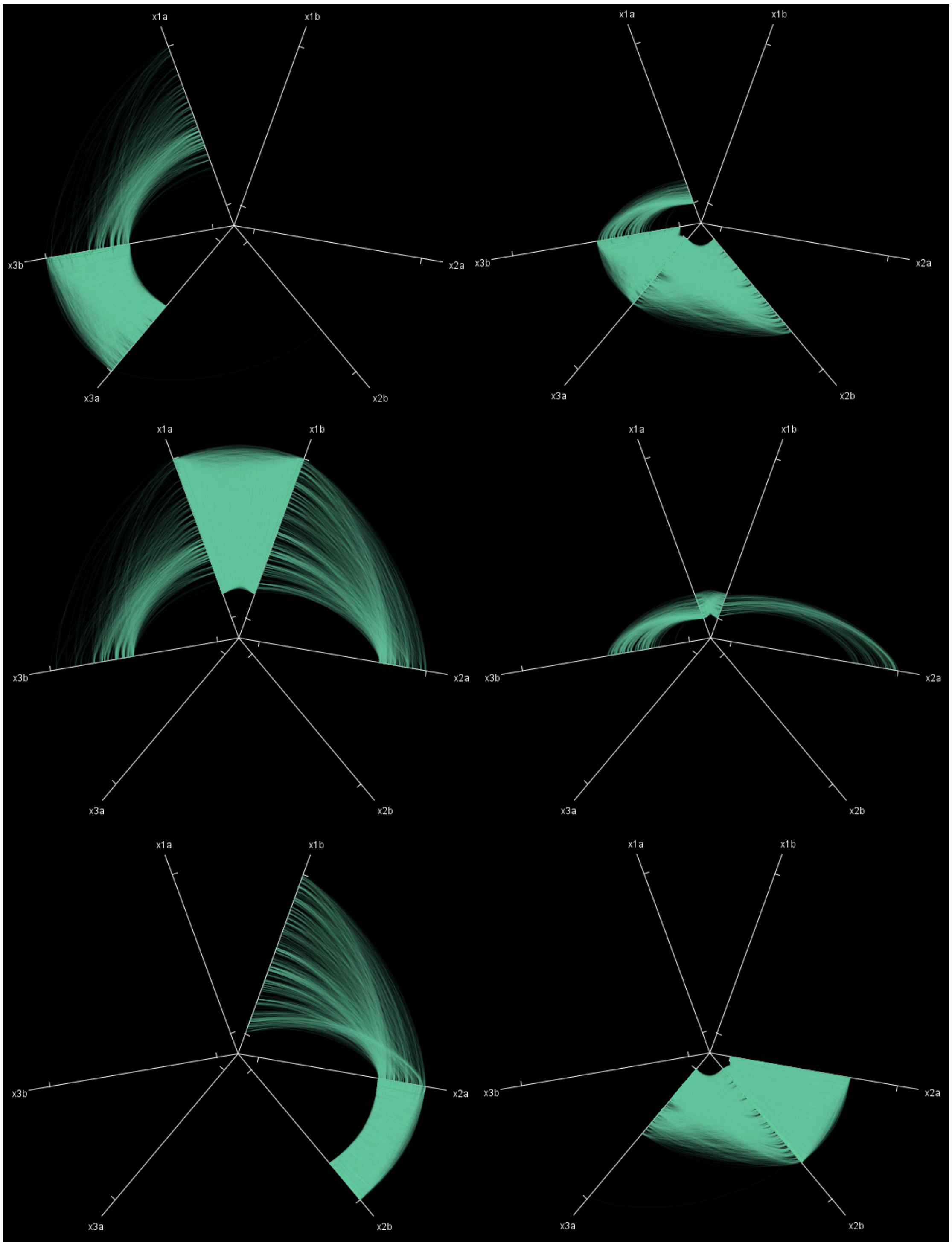


FIGURE 5.5: The figure shows all the detected circadian modules. Each axis depcited in the Figure is a split axis showing one of the modules detected via ClusterONE. Each axis is furthermore split into 2 sections. The connections are thesholded and show the intra module interactions versus the inner module interactions. Genes only having connections with others in their own cluster, or with the two neighbouring clusters are visible.

Statistics	**module1**	**module2**	**module3**	**module4**	**module5**	**module6**
Number of nodes	544	436	213	237	200	86
Number of edges	73780	45242	11333	13694	8947	1474
Average Degree	271.25	207.53	106.41	115.56	89.47	34.28
Density	0.5	0.47	0.50	0.49	0.45	0.40
Modularity	0.149	0.167	0.134	0.187	0.169	0.203
Average Clustering Coefficient	0.708	0.711	0.713	0.719	0.700	0.661
Eigenvector centrality	0.00349	0.00363	0.00211	0.00295	0.00383	0.00141
Average path length	1.500	1.525	1.499	1.513	1.559	1.616

Table 5.1: The table portrays topological properties of the six modules identified using ClusterONE. Modules five and six are subdivisions of the four eigenmodules detected previously, hence the decreased number of edges and concurrently average degree is visible.

with the neighboring modules instead of opposing modules. This makes sense in evolutionary terms where the modules demarcated compact processes and the interactions between these have a smaller number of connections. The table supplements the module information by network characteristics.

5.2.3 MODULE COMPARISON

There are multiple methods for module comparison particularly for comparisons across different data sets. The matrices used to detect modules in the previous sections (Chapter 3, eigenanalysis) and the obtained modules themselves are being explored in this section. Foremost it is worth mentioning that modules had to be inspected across the same experimental design and then across closely related species. Module preservation statistics can be used to evaluate whether a given module defined in the reference dataset can be detected in the test set. For correlation networks methods include: correlations of correlations, correlations of eigengene-based connectivity as some of the options that can be used. There are two fundamental questions to be addressed. Are the modules (as groups of genes) denser than background? Is hub gene status preserved between reference and test networks? The former question identifies network motifs (statistically significant subgraphs of a larger network). The latter reflects on the evolutionary origins of edges in such a network. These two questions were answered form the perspective of the circadian modules. This is of biologial relevance as the motifs might reflect old and novel loops both essential and redundant whereas the hub status may help determine the evolution of the networks hence the core loops/ To obtain clear and unbiased results the partitions obtained from the previously described module search served as input. There are numerous methods to compare partitions and these include pair counting (geometric mean), Cluster matching (RAND index, Jaccard score), Information theory (normalized mutual information) and Graphical Gaussian Models to name the dominant trends. The module comparison was attained using three of the dominant measures, these being the

Jaccard index, the normalized mutual information and the adjusted RAND coefficient. The methodology is described in Chapter 2 and here are the operating principles

- (Nr of edges shared among 2 sets)/(Nr of edges shared among 2 sets) + (specific for A) + (specific B)

- Difference between RAND is in (*) the number of pairs of nodes that are in different communities in both partitions added to numerator and denominator

- 1000 realizations

- the similarity of two modules is measured by the Jaccard index score between the edges of two co-expression graphs whose nodes are the members of the modules and whose edges are those pairs with a co-expression >= 0.8 (binary attributes).

These results for network motifs comparisons in Arabidopsis and other species were depicted in the Table 5.2 and Appendix. The results are indicative of strong preservation (produced a normalized similarity score for each subcluster expressed as the number of standard deviations from the mean of the distribution of Jaccard similarity scores for equivalent randomized module structures). Additionally, the results for interspecies and interdata module comparisons for selected genes were inspected using the '916 project' encompassing 822 experiments that is all the Arabidopsis experiment deposited in GEO at the time of the study used to form the background set. It was incredibly important to see the agreement as such a grand experiment certainly does provide greater confidence. For the four main circadian modules the Jaccard scores are presented in Chapter 3, specifically the overlap between phase module compositions in Arabidopsis thaliana and Oryza sativa (with randomizations used to assess the significance of each score). They do show preservation hence are likely resembling the core loops as they are consitent across datasets coming from the same experimental design. Furthermore the details are presented in Table 5.2 where the empirically derived random model was used to assess the distance Jaccard scores. Such preservation is without doubt not coincidental.

5.2.4 MODULE ENRICHMENT

Once the networks motifs were identified and verified in the previous sections. Here it was interesting to ask in terms of biological relevance which process's pathways they were enriched for and which of these were 'disturbed'. Module functional enrichment was carried out using the same Arabidopsis circadian gene expression datasets and the

ID1	ID2	STD	Mean	Jaccard	Normalized x STD
class1	subcluster1	0.0036	0.0399	0.2622	61.60
class1	subcluster2	0.0039	0.0429	0.2811	60.42
class2	subcluster3	0.0069	0.0208	0.2657	35.00
class3	subcluster4	0.0013	0.0881	0.1935	76.43
class3	subcluster5	0.0018	0.0796	0.1605	44.46
class3	subcluster6	0.0021	0.0872	0.1734	39.75
class4	subcluster7	0.0022	0.0714	0.2243	68.05
class4	subcluster8	0.0023	0.0658	0.2250	65.99
class4	subcluster9	0.0026	0.0672	0.2123	46.07
class4	subcluster10	0.0031	0.0672	0.2123	46.32
class4	subcluster11	0.0037	0.0548	0.1548	26.49
class4	subcluster12	0.0044	0.0550	0.1619	23.95
class4	subcluster13	0.0046	0.0574	0.1652	23.39

Table 5.2: The table depicts the subclusters which were inspected in terms of preservation in Arabidopsis and across species. The number of standard deviations away from mean reflects on the true existence of the network as in this example it was inspected against randomly shuffled network (preserved degree).

mutant *far1 far1fhy3* dataset, performing both experiments from the perspective of the nodes and the edges. In this case, the decicion was to expand the analysis to look at other species. Node GO enrichment is a simple hypergeometric over representation which was performed. The expression data for soybean (15,753 genes), poplar (28,969), grapevine (8,255 genes), rice (34,153 genes), and maize (10,068 genes) were assembled from the NCBI Gene Expression Omnibus. Edge enrichment involves a differential analysis, essentially between changing conditions and is outlined below.

1. Take all the nodes (genes) in a particular GO category.
2. Take all the edges (interactions $->$ correlation >=0.8) between all the nodes in particular GO category in WT.
3. Check how many edges found in A (corr>=0.8) appear in B and C (corr>=0.8).
4. The higher the percentage between A-B and A-C the less disturbed GO module due to the mutational status.

The same principle was applied to the *far1* and *far1fhy3* mutant dataset. The same differential methodology is applicable to detect the GO disturbed categories when mutants are compared, hence the procedure described below reflects exacly that.

1. Take all the nodes (genes) in a particular GO category.

2. Take all the edges (interactions $->$ correlation >=0.8) between all the nodes in particular GO category in WT.

3. Check how many edges found in WT (corr>=0.8) appear in *fhy3* and *fhy3far* (corr>=0.8).

4. The higher the % between WT-*fhy3* and WT-*fhy3far* the less disturbed GO module due to mutational status.

That led to the analysis of the most disturbed processes (adhering to the methodology outlined above) further discussed. In terms of simple enrichments all the mentioned and discussed modules were analysed using DAVID (Huang et al., 2009). Among the most disturbed processes is 'GO:0009765', photosynthetsis and light harvesting, 'GO:0009595' detection of biotic stimulus, 'GO:0045087' 'GO:0009867' immune response and jasmonic acid mediated signalling pathway. The results were highly interesting in terms of the mutants where the most disturbed process between WT and the single mutant (*fhy3*) was GO:0005983 the starch breakdown (96%) and senescence among other interesting terms (83%).,In terms of the double mutant and the disturbed processes, GO:0010206 photosystem II repair was disturbed, as was GO:0009765 photosynthesis light harvesting. Among processes disturbed in both mutants was GO:0000373 being Group II intron splicing.

5.2.5 LINKING PHENOTYPIC TO MOLECULAR NETWORKS

Further analysis was then carried out with the aim to link standard circadian gene coexpression networks described above and PPI networks derived from the same pool of proteins in Arabidopsis. Co-expression networks carry a great deal of information. Nevertheless *in silico* analysis often times greatly benefits from validation and one such form of validation is an overlay of one network with another one formed from the same components but containing another set of information, for example protein-protein interaction data. The PPI information may be revealing in terms of shared functions, but it does not indicate hierarchy of regulation. The reason it has been used is a. to find out the extent of agreement and b. to verify with certain known circadian PPI that have been already annotated (in order to see new ones). A connection between two genes in a coexpression network does not have to correspond with a connection in PPI networks, pathway and regulatory network. It is important not to confuse the edges of gene co-expression networks as direct physical interactions. Assuming that these connections lead to finer levels of granularity like for example pathways, these would still have to be confirmed experimentally. It has been previously exemplified in yeast that topological comparisons indicate that co-expression networks are not directly related to the PPIs.

The Arabidopsis Interactome Mapping Consortium has mapped roughly 6,200 reliable interactions between about 2,700 proteins using the yeast two-hybrid-based system. This data was used to compare the distribution of correlation between all possible interactions defined by the PPI network of the 196 circadian genes (all present in the dataset) and an equal number of random correlations in Arabidopsis *thaliana* (PPI vs coexpression) (PPI vs random). That is depicted in Chapter 3 in Figure 9. Comparison of the distribution of correlation was performed between all possible interactions of 196 genes and a selected equal number of random correlations in Arabidopsis *thaliana* (PPI vs coexpression) (PPI vs random). The 196 circadian genes present gave rise to 37 interesting

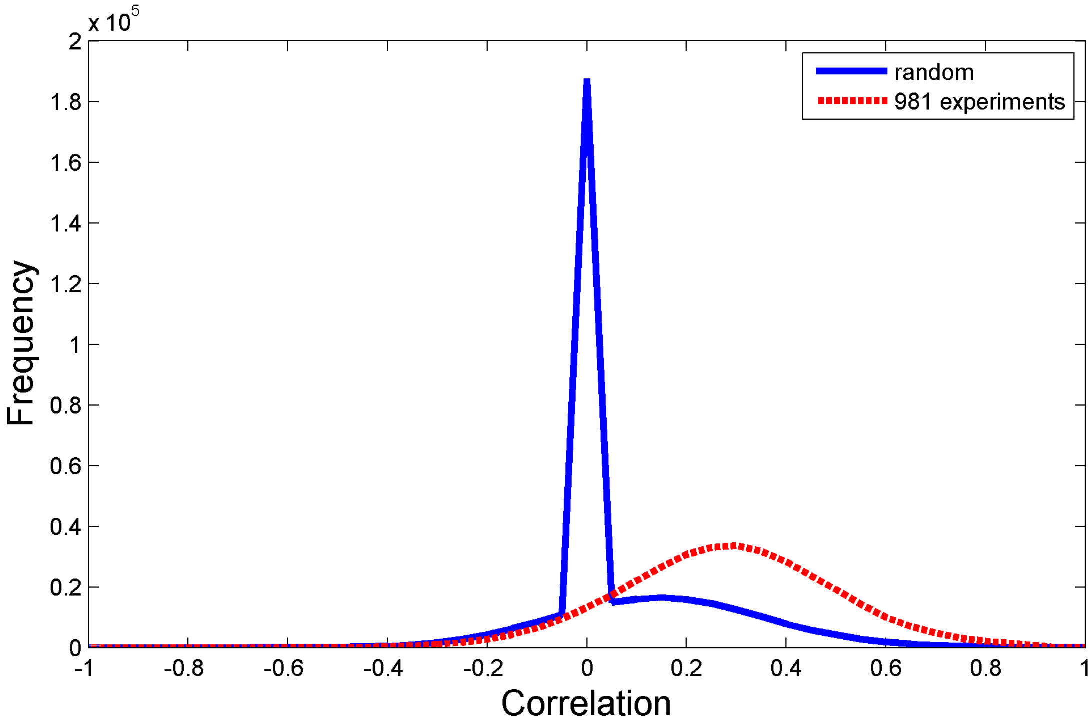


Figure 5.6: Two probability density functions for correlations between PPI vs. coexpression and PPI vs random set were compared with blue line representing random model and red line representing thresholded gene coexpression gene.

interactions according to a molecular network PPI. Figure 5.6 portrays the overlay of circadian static correlation dataset with the large 961 compendium encompassing over 4000 microarrays. The compendium datasets contain all the Arabidopsis experiments which served as a coexpression set for all the genes. Certainly there is greater agreement between the PPI and gene coexpression than random as was expected.

Another experiment simply integrated BIOGRID PPI data to overlay with a circadian static experiment yielding validating results and this is portrayed in Figure 5.7. We learnt two things: there is a degree of agreement between coexpression and PPI on the inspected gene pairs and secondly we have a range of feed forward loops and motif types that are highly interesting and never seen before (to the best of my knowledge).

Figure 5.7 displays these overlaid genes, detected in both datasets. That shows interesting interactions like those between *PIF4* and *PHYA* and *HFR1*. The example nodes that are coexpressed and interacting. Secondly this is a proof of concept in terms of joining multiple network types. The major difference between such a comprehensive PPI dataset obtained from one set of experiments and BIOGRID is that the latter collects data from many studies.

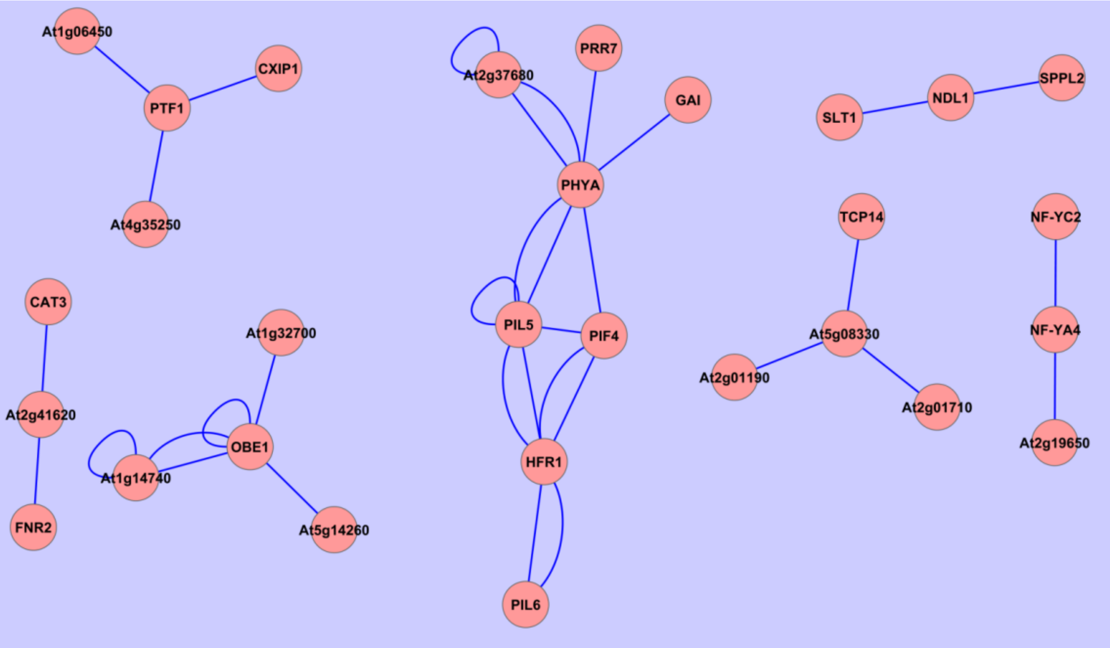


Figure 5.7: Shows circadian PPI interactions overlaid with circadian coexpression network derived interactions. It shows seven distinct modules with several autoregulatory edges.

5.2.6 DIFFERENTIAL INTERACTION MAPPING

We then aimed to carry out differential network analysis on the same circadian dataset yet focusing on day and night separately and using the *far1fhy3* and *far1* mutant dataset. In the differential network analysis several design options were considered. We computed the 'difference score' that quantifies the change of genetic interaction across two conditions (2 experimental designs). The maps obtained after subtraction were compared to static maps (leaving the differences). Furthermore the objective was to inspect the obtained maps for presence of already discovered and described modules. These were already inspected for enrichment of particular GO categories (Chapter 5, module enrichment). Findings suggested that differential interaction networks may reveal the processes that are dynamically engaged during cellular responses to for example external conditions. This led to the hypothesis tested in the next sections, specifically, differential functional networks delineate the intricacies of regulation of the circadian clock. If genetic interaction maps are the result of the context-dependent wiring of regulatory networks in the cell, this suggests that changes in one layer are reflected in the other and vice versa, opening a number of exciting and interesting possibilities. For instance, if dynamic changes in the topology of molecular interactions could be used to predict the corresponding changes in genetic interactions, this would open exciting possibilities. That will be evaluated in the chapter 6. The aim in this experiment was to construct a global map of modules and their dynamic interactions across two temporal conditions. The inter module interactions were quantified. That was further inspected using specific TF, which reflects the rules that govern specific changes in topology that are likely causative instead of simply descriptive.

The time day and night experiment Sampling for the first circadian time series microarray: 0h -day/night hence excluded
4h - day
8h - day
12h -day/night hence excluded
16h - night
20h - night
24h -day/night hence excluded
28h - day
32h - day
36h -day/night hence excluded
40h - night
44h - night
The second dataset was the *WT/fhy3/fhy3far1* mutant dataset with 4 timepoints.

Three experiments were performed to assess selection criteria. Differential interactions were assessed by computing the difference across 2 periods. The null distribution was established (by comparing two consecutive days). The differential interaction p-value was established. If correlation in condition 1 is higher than that of condition 2, it was considered ”positive differential” and if interaction for which condition 1 is less than condition 2, it was considered ”negative differential” (refers to Figure 5.13). For example, if one imposes such a method, a cutoff value obtained from a null model (day1 expression vs day2), with $p<0.0001$ is found at Correlationday - Correlationnight > 1.7134 for ”day” specific edges and at Correlationday -Correlationnight < -1.7168 for ”night” specific edges. Three approaches were tested here before the results were selected. The second approach was selection for high positive correlation in day condition versus negative correlation in B condition. The third approach tested set criteria for high positive correlation in the day condition versus no correlation (between - and + 0.2) within the night condition. The principles 2 and 3 are portrayed in Figures 5.8 and 5.9

whereas first network derived using method 1 is presented in Figures 5.13. Specifically Figures 5.8 and 5.9 trace a gene linked to CCA1 given 2 set of aforementioned criteria. In Figure 5.8 the first gene is maybe influenced by other non-circadian factors at night while the latter is a double-peaked gene, possibly influenced by a second evening phased circadian influence as well as the morning phased circadian influence that links it to CCA1. As the high versus no correlation method was selected as optimal, selected genes were

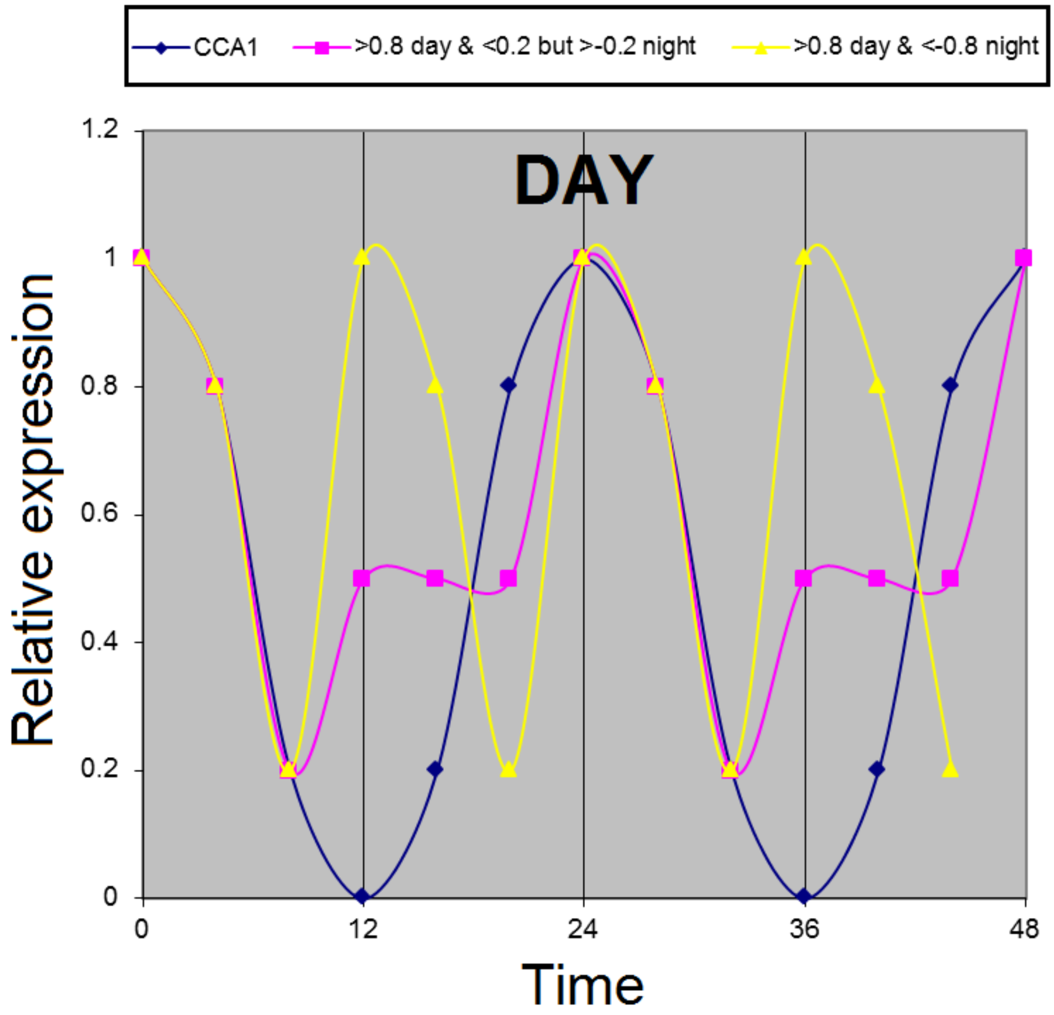


Figure 5.8: Potential variations in the design of differential interactions experiments. The method annotated in pink detects the strongest correlations in a given condition and low (0.2/-0.2) in the other condition. This is in contrast to the method which detects high positive and negative correlations (yellow). In the time circadian inference the greater emphasis will be placed on the pink precondition.

mapped accordingly with their day and night specific groupings. This was preferred in favour of a switch to a negative correlation which would call for a unreasonable pattern much less likely than a simple loss of correlation if something else were to become a controlling factor. This result is represented in Figures 5.10 and 5.11. Figure 5.10 depicts gene expression linegraphs of selected circadian genes and their differential interactors

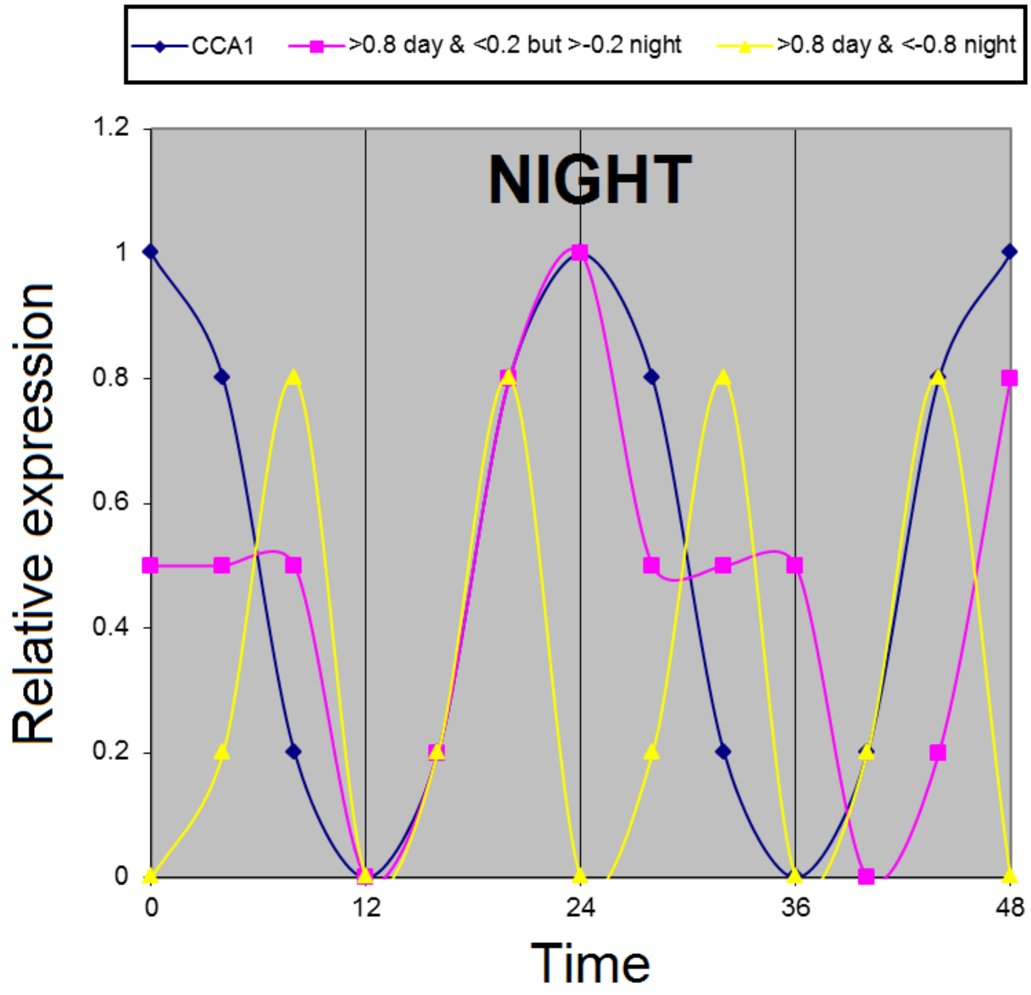


FIGURE 5.9: The possibilities of designing the differential interaction experiments. The method annotated in pink aims at detecting the strongest correlations in a given condition and low in the other versus high positive and negative correlations (yellow). In the time circadian inference the greater emphasis will be placed on the pink precondition.

during the day and night. At the same time it presents the potential reordering of the timepoints given the resolution with the aim of inference of day and night events. This approach also demonstrates wider possibilities as with greater granularity more evidence should be available on, for example central hubs and their rearranging partners. Figure 5.11 displays the concepts of differential interactors. Essentially, whereas Figure 5.10 displays the entire profile, 5.11 displays cut data with the interactors captured relying on PIF4. The probability density functions (PDF) to justify the selected method are graphed on Figure 5.12 where there is a clear boundary between the random and the selected method. The hypothesis here is that the PDF will show whether there is likely to be some genuine difference between the PDF for all the time points (blue) and the PDF for the selected data (green). If there is a greater difference between the green and the red versus the blue and the green then it is reassuring that even though there are a

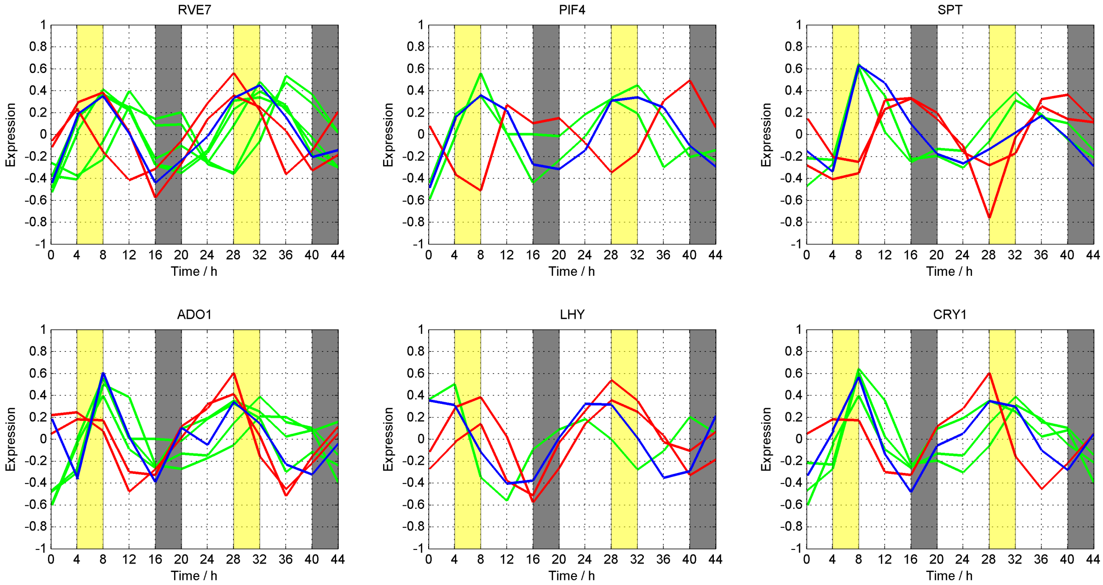


FIGURE 5.10: The gene expression line graphs of the selected genes and their differential interactors during the day and night. The day and night specific interactors are portrayed in green and red respectively. The core gene is annotated in blue.

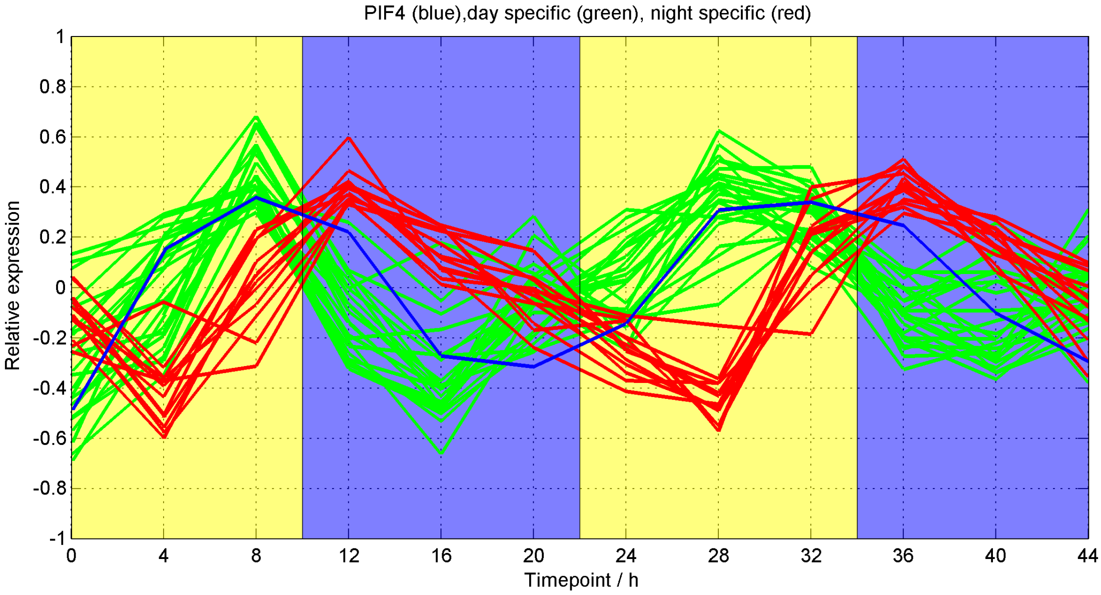


FIGURE 5.11: The concept of differential interactors where the time vector has been altered due to limited number of timepoints. The core gene is blue, the day interactors are shown in green whereas the night interactors are shown in red. The vector of expression has been altered showing consecutive day and night points.

few timepoints they can regardlessly be used in this study. Essentially, the probability density function (PDF) of the entire vector versus combined vector and a random vector were used to assess the extent of differences in order to determine if such approach is possible.

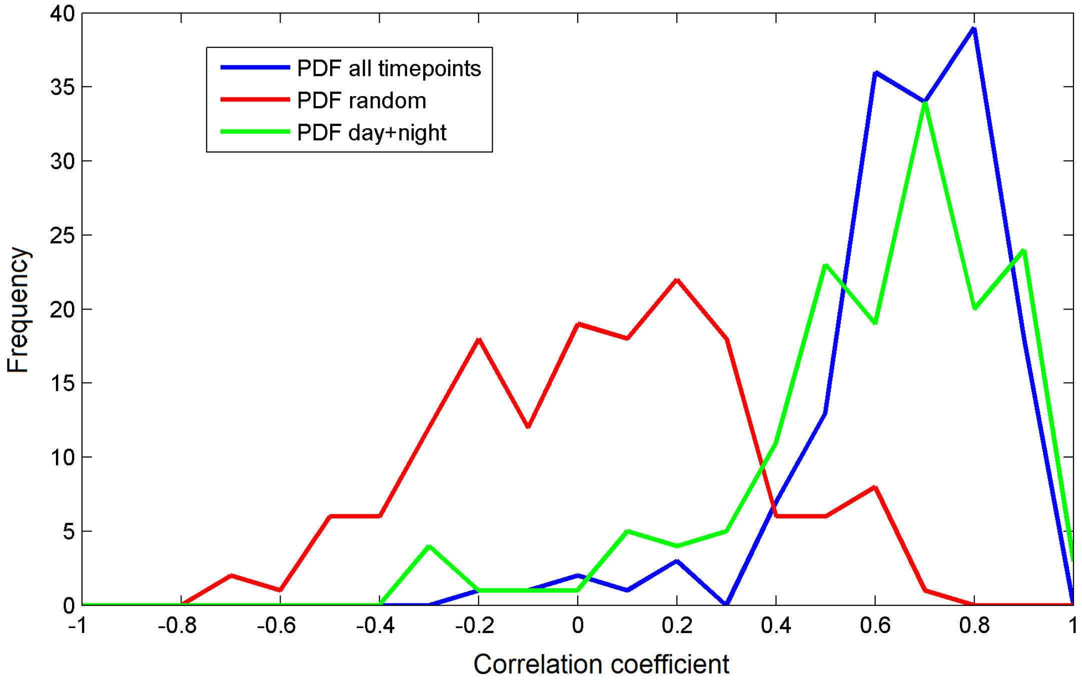


Figure 5.12: The proof of concept evidence for the use of day and night and all timepoints in the differential interaction analysis (x axis represents correlation, y axis represents frequency). The blue line represents probability density function for all the time points. The green line represents the PDF for the selected data whereas the red line represents the randomly permuted data.

The initial inspection of the series of methods yielded highly interesting results. Among 1478 selected genes based on their pairwise differential correlation (high during the day, less so at night), there are solely 12 night interactions where the number of day connections is 22581 (interactions that happen in the night points only which are not present in the day points). It is important to stress that this is not by any means a foregone conclusion. Genes showing the differential correlation pattern in question in a pairwise manner are not restricted on any particular expression pattern so the fact individual genes, selected because they are part of pairwise patterns correlated based on the day part of the cycle, are so rarely found to correlate based on the night part of the cycle is highly unexpected and likely to be representative of biological significance. This is demonstrated in Figure 5.13. A similar situation is observed for selected genes based on their pairwise high correlation during the night and lack of correlation during the day.

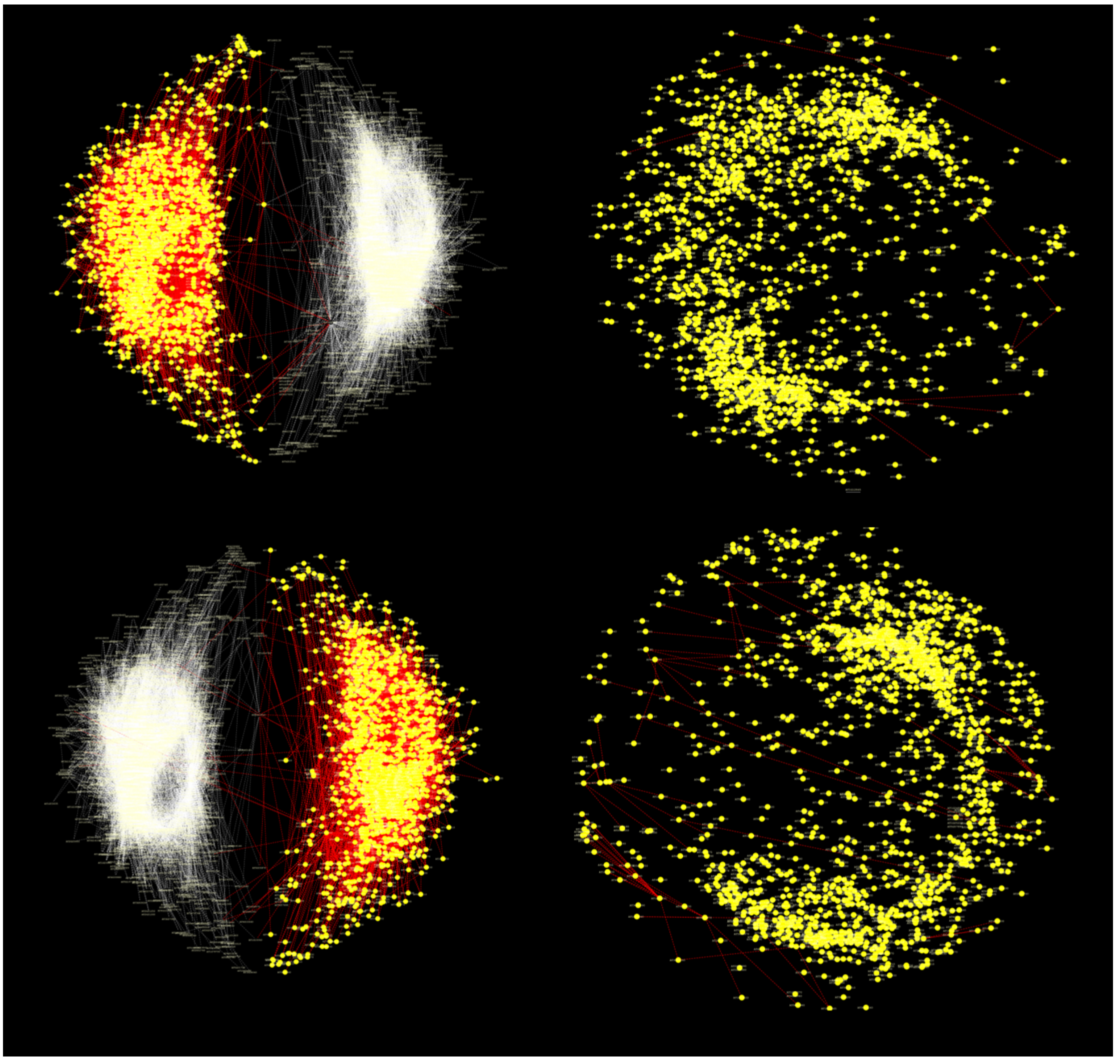


FIGURE 5.13: The day specific and night specific interactions obtained through differential interaction mapping, using method 1 (If correlation during the day is higher than that of night, it is considered "positive differential" and if correlation is less than night, it is considered "negative differential"). The left panel portrays day (upper left) and night (lower left) specific intereactions whereas the right panel portrays those day and night nodes during night and day respectively (and their respective edges in red colour).

Figures 5.14 and Figure 5.15 map the same global and time specific layout. These day and night networks were produced through alternative criteria, selection for high (0.8 and greater) positive correlation in the first condition versus negative correlation in the second condition. Both methods yield resemblance particularly in terms of the 2 clusters forming in the day and night network. There is a greater number of interactions at night between the 2 observed clusters. Furthermore, apart from the global analysis, individual genes were selected for including *ELF4*, *PIF4* and *RVE8*. For example *ELF4* expectedly has multiple members during the night and not so many during the day (a 1:19 difference of specific connections). These example trials served as the foundation for the differential analysis outlined in the next section.

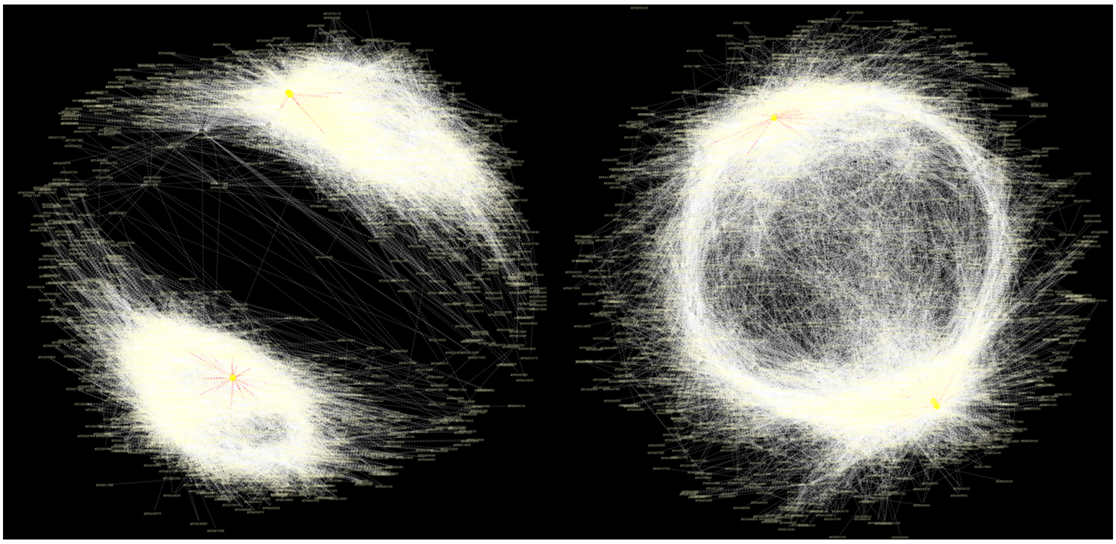


Figure 5.14: The day and night networks were produced through selection for high positive correlation in the first condition versus negative correlation in the second condition. The yellow genes are *CCA1*, *(*LHY) and *TOC1* for orientation purposes and edges are marked in red. The left panel is day network and the right panel is night network

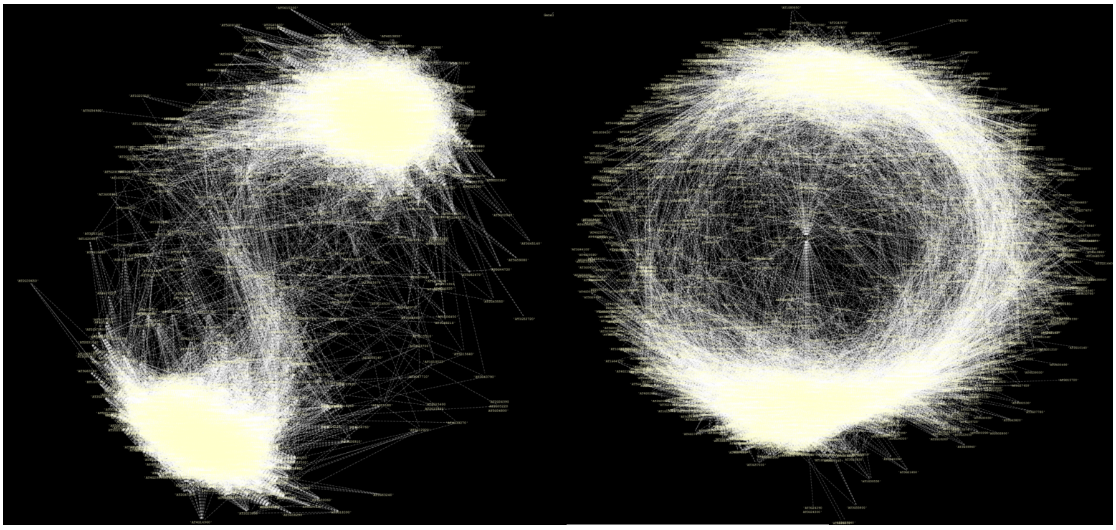


Figure 5.15: The third approach tested criteria set for detection of a high positive correlation in the day condition versus no correlation (between - and + 0.2) within the night condition. This was the optimal method of choice for selection of differential interactions. The left panel is representative of day and the right panel is representative of the night network

	Day	**Night**
nodes	2860	2877
edges	11395	6041
density	0.003	0.001
avg. degree	7.969	4.2
clustering coef.	0.534	0.492
PCA	3	3

Table 5.3: Topological properties of day and night networks.

5.2.7 REWIRING OF NETWORKS ACROSS TIME

We constructed the individual networks to examine the specific pattern of associations exhibited under each condition (day and night). These were constructed using At coexpression network data. The construction of the individual networks relies upon a binary existence of correlations (significant and non-significant) hence the correlation strength is not represented. The left panel of Figure 5.16 indicates the static pattern of pair-wise association for day selected edges. The night network (right panel of Figure 5.16) reflects the pattern of associations for night activity. The aim was to explore variations in the observed pattern of associations between the two time-groups and to delineate some changes in the features of such networks. The patterns of connectivity for each of the nodes can be used to define the major differences between those two time periods. Despite the extensive similarities between the two networks, some distinction in the pattern of association between time frames can be identified. The day network is more dense and there seem to be some hub like paths in the centre. Despite the same almost identical number of nodes 2860 and 2877 respectively, the number of connections is by far reduced at night from 11500 to 6050. The global properties of these networks are given in Table 5.3.

The networks are similar with clear difference in activity, with the night graph being by far more sparse. The clustering coefficient is another measure used to describe the organization of complex networks. It has been suggested that high value of this measure can be an indicative of the modular organization of the network. The mean clustering values for the day network and the night network are 0.53 and 0.49 respectively. A central property of scale free networks is the existence of a relatively small number of highly correlated nodes, the circadian hubs. These highly connected genes are central components of the networks reflecting focal points of the network organization. These are depicted in Figure 5.17 where the betweenness centrality is presented in relation to

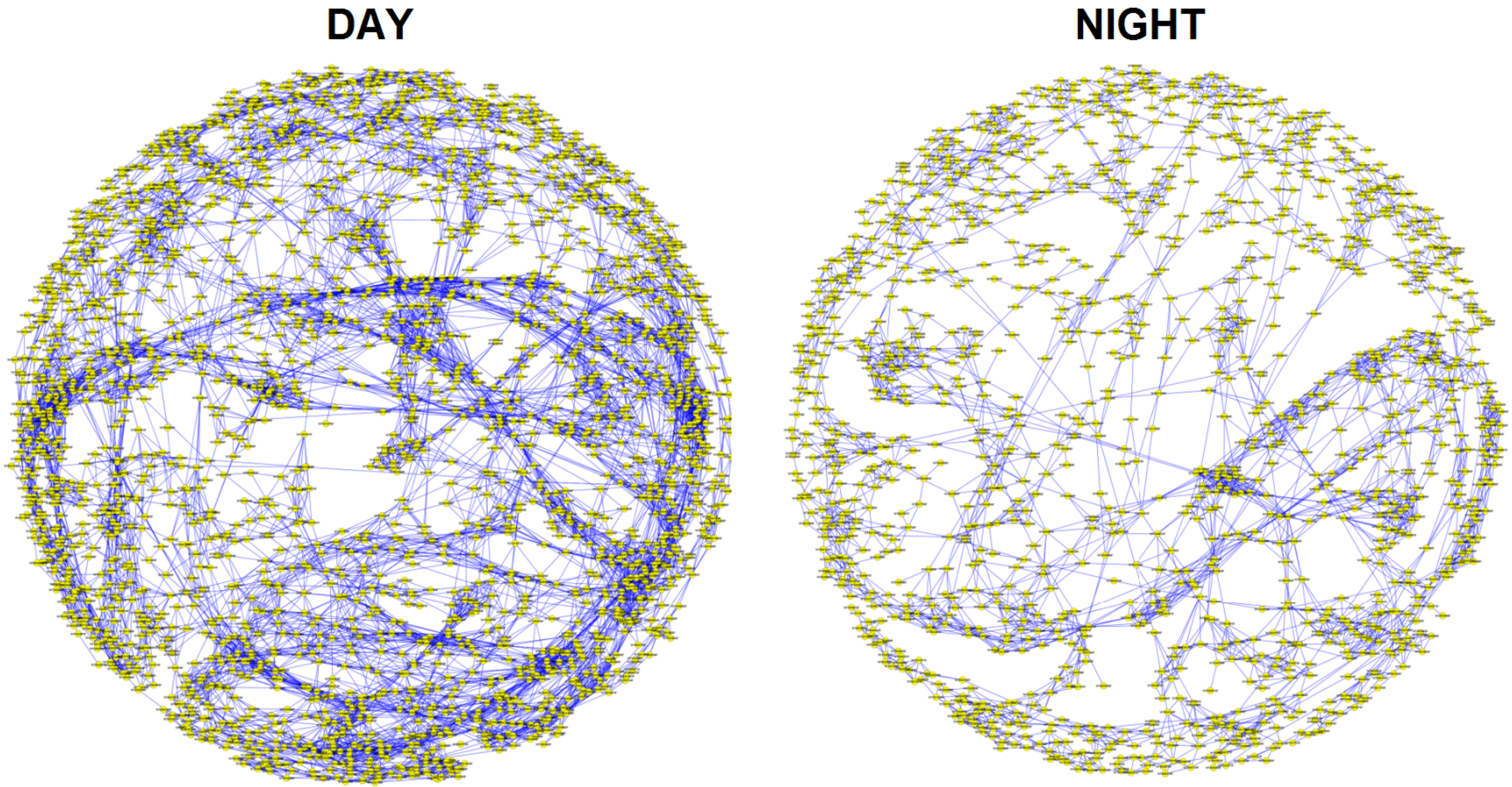


FIGURE 5.16: Individual static networks where each of the edges indicates a significant pairwise correlation between the genes using a corrected threshold $p < 0.01/z$, where $z = M(M-1)/2$ and M is the number of genes. Each network is derived on the basis of 6 timepoints, 6 per day for the day network and 6 per night for the night network. It is the global topology, pattern and density that are important features of these networks, not the individual links, hence the size of the nodes.

the degree connectivity. The outermost genes represents central nodes in the network. Essentially the higher the degree connectivity the more central the node is. When looking through the betweenness lists there are some interesting hits present, these being NPX1 gene (previously unreported hub) central to the day network and the *RVE8* (AT3G09600) as the central night hub (in agreement with Hsu et al. (2013)). *RVE8* is a MYB family transcription factor, when overexpressed resulting long period phenotype. It functions in a positive loop with *PRR5* and binds to the EE in *TOC1* promoting its expression. This could be highlighting the output links between the clock and the rest of the genome reflecting how output links change with time.

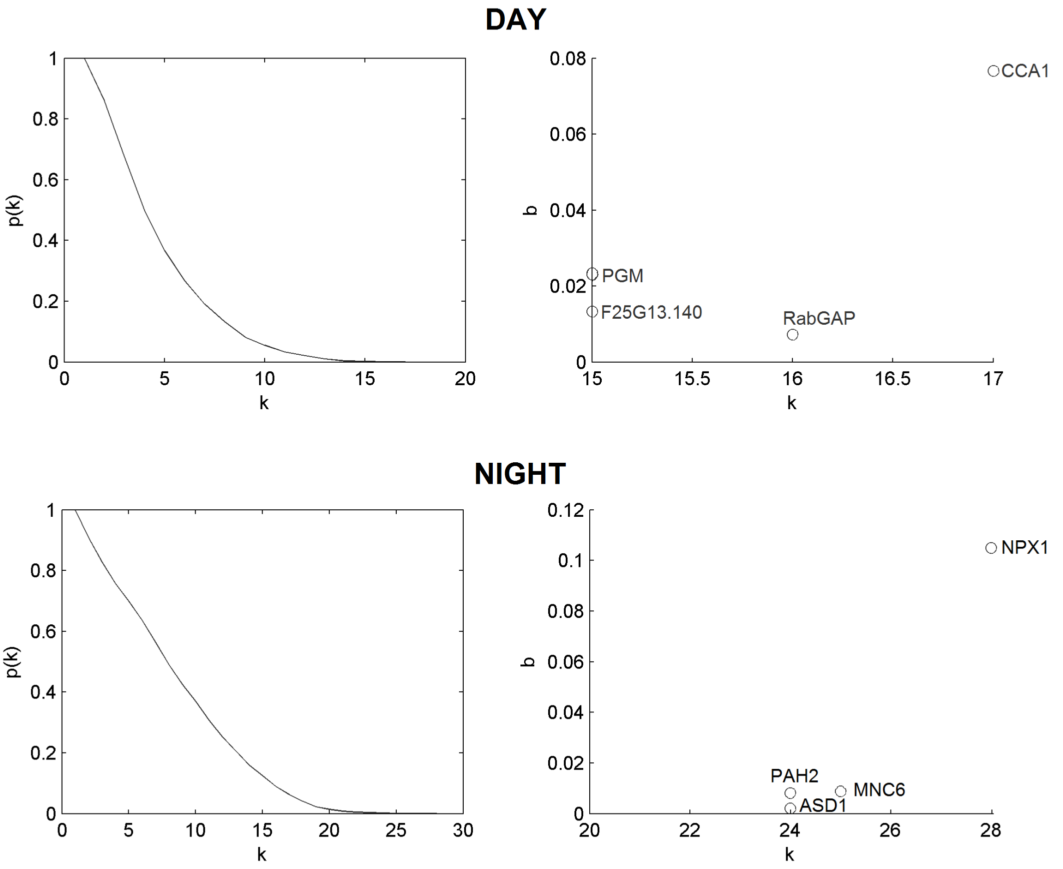


Figure 5.17: Topological properties of the individual networks. The degree distribution shows the probability that a node i has k connections in the network. For both individual networks the degree distribution is well approximated by a power-law. The betweenness plots on the right side of the panel indicate how central a node is in the network. For both individual networks, the nodes with higher degree connectivity are also the most central nodes.

To explore the functional differences between the two groups a two sample Mann-Whitney test was performed. Initially the pattern of associations was investigated for each group, secondly, the test was carried out comparing the two networks with a precondition being correlation greater than 0.8 and between -0.2 and 0.2 in the opposite condition. We define edges by computing *(*p) values for the two sides test with the null hypothesis that the correlation of two variables is the same versus the alternative hypothesis: the correlation of two groups is different. The test allows one to reject the null hypothesis. The differential network inferred is presented in Figure 5.18.

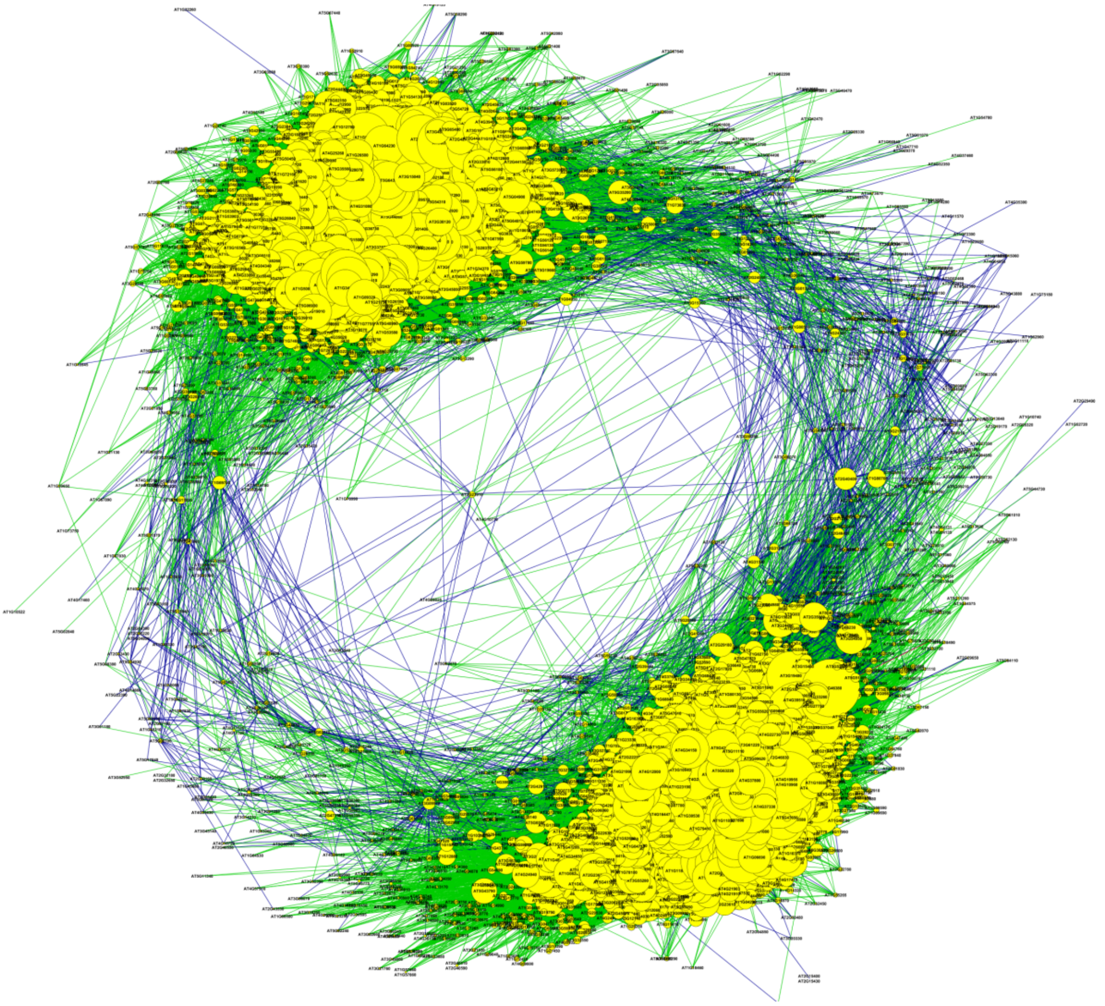


FIGURE 5.18: The differential network obtained from day and night using the 0.8 versus -0.2 - 0.2 approach method 3. Each connection indicates a significant change of the association measure across the changing conditions. The colour depends upon the direction of change (green = day correlation) (blue = night correlation).It is the global topology, pattern and density that are important features of these networks, not the individual links, hence the size of the nodes. The split day and night network is visible in Figure 5.19

The individual graphs are depicted further on Figure 5.19. They are essentially representing the split version of Figure 5.18. There are 2 apparent clusters visible during the day which are gone during the night. One is right away puzzled by the sparseness on the night graph, by the 2 clusters forming during the day and by the circular arrangement of the night activity. As with the individual networks, topological properties can be

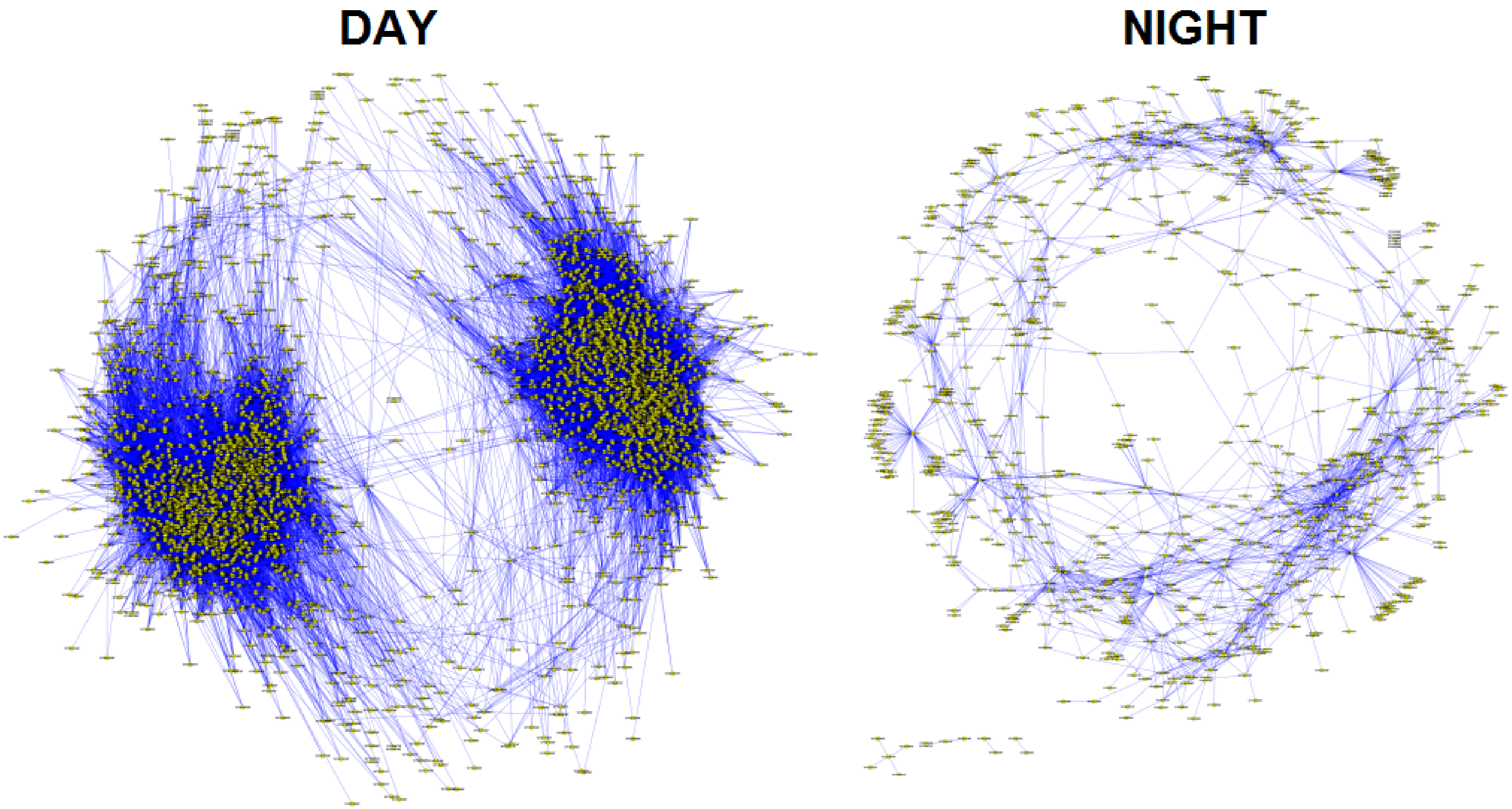


FIGURE 5.19: The distinction between the obtained day and night differential interactions quantities with day being the left graph and night being the right graph. The night graph is sparse and the interactions (correlation) are dispersed in comparison to the day network.It is the global topology, pattern and density that are important features of these networks, not the individual links, these are evaluated in the text.

used to describe the differential networks. The hubs for these networks are presented in Table 5.4.

The next sections of the analysis was fundamental to the hypothesis being tested, what modules, processes are most disturbed between the networks (more explanations in the methods section). The answer here is photosynthesis amongst several others. This process is 'GO:0019216' a lipid metabolic process. This is proceeded by 'GO:0009765' photosynthesis and light harvesting, 'GO:0009867' the jasmonic acid signaling pathway and 'GO:0045087' immune responses. The essence of such networks can be further evaluated by looking at Figure 5.20 were the genes in the GO categories, core clock and clock regulated genes were mapped in the same way. Here of great interest are *PIL4* having several night collaborators, *RVE7* which seems to be the connector between day and night and the *MPK7* which clearly plays a hub role in the day network. One highly interesting result that comes out of the recent differential analysis is the hub gene EARLY-PHYTOCHROME-RESPONSIVE (*ERP1*, *RVE7*) MYB family transcription factor. It is located between *PIF4* and *PIF5* and has a series of day specific and night specific interactors. It has a night specific edge with *PIF5* and a day specific edge with *PIF4*. Furthermore it has a day specific link with *PRR7*. *EPR1* comes into physical interaction with *TPR1*, *TPR2*, *TPR3* according to BioGRID it is regulated by phytochrome A and phytoschrome B. It was shown to control *CAB2* expression and furthermore displays autoregulation (Kuno et al., 2003). Results indicate that *EPR1* overexpression causes a slightly altered photoperiod flowering response. Furthermore the increased levels of *EPR1* affect several phytochrome and circadian rhythm regulated processes. In continuous light it has a peak expression levels occurring at subjective dawn and has 3 *CCA1* recognition motifs. As Figure 5.21 clearly portrays the difference between static and differential interactions reflected in the intra and intermodular behaviors in the circadian system per se. There are more differential interactions between modules showing that these processes are established and that their interplay is the core modus operandi, architecture of the clock.

5.2.8 REWIRING OF NETWORKS ACROSS MUTANTS, GENOTYPIC DIFFERENTIAL INTERACTIONS

These networks investigated here were using the microarray experiment on *far1 fhy3* single and double mutants, grown in 12-hour light 12-hour dark rhythm at 21 degC for 1 week, transferred to continuous red light, samples taken at 8 hour intervals, 4 time points. The principles of the statistical analysis resemble those used for the time experiment. Here the 'conditions' are of interest rather than the time notion. The focus is on differential networks yet a list of differentially expressed genes is present

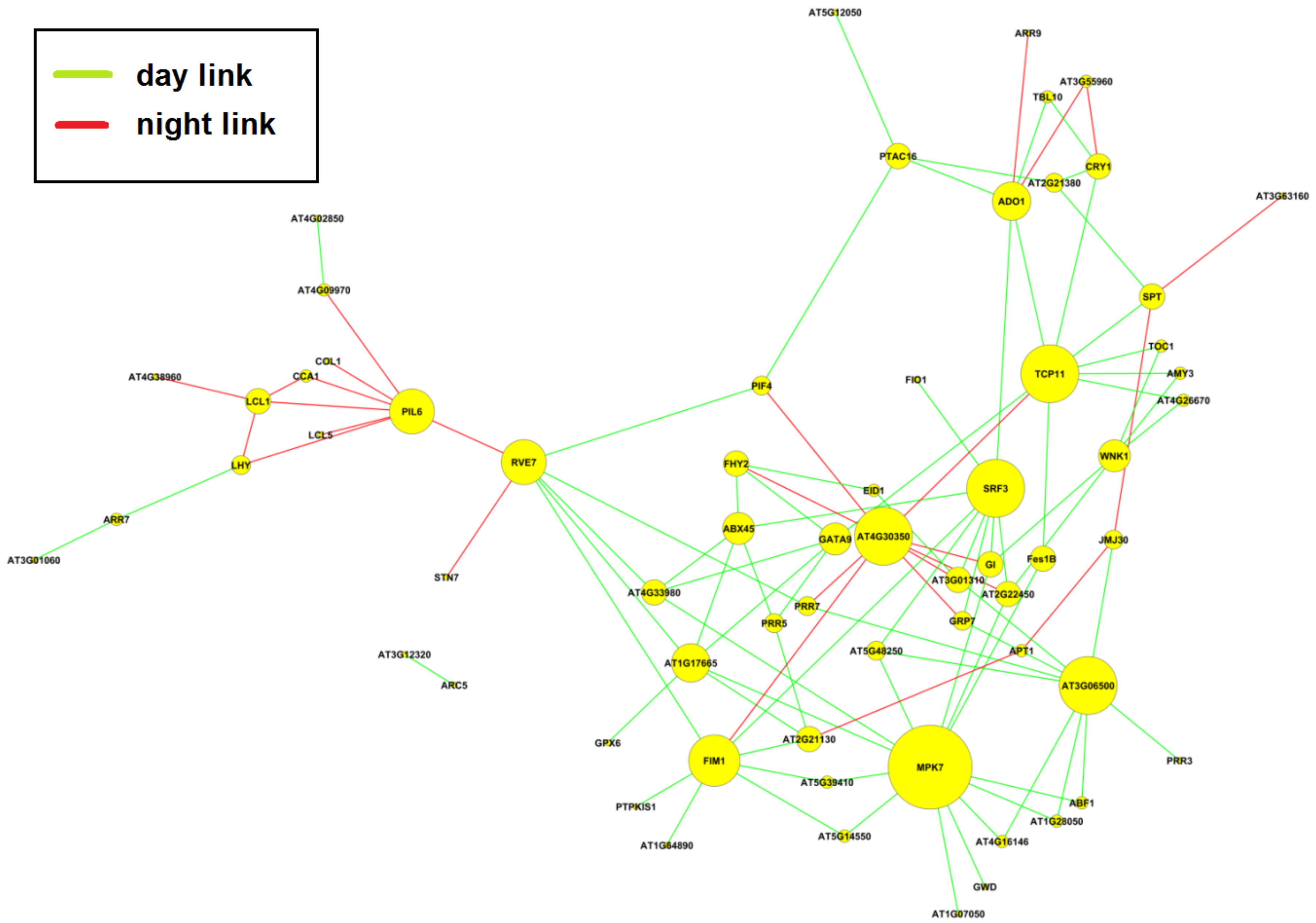


Figure 5.20: Shows the central hubs of the differential day and night networks. Here, clearly the important role if *RVE7* is seen as the hub connecting day and night modules. The paths in and out degree is essentially the bottleneck in the present network. Additionally both *PIL6* and *MPK7* come across as central hubs as does *TCP11*. *RVE7* will be further investigated in the next chapter.

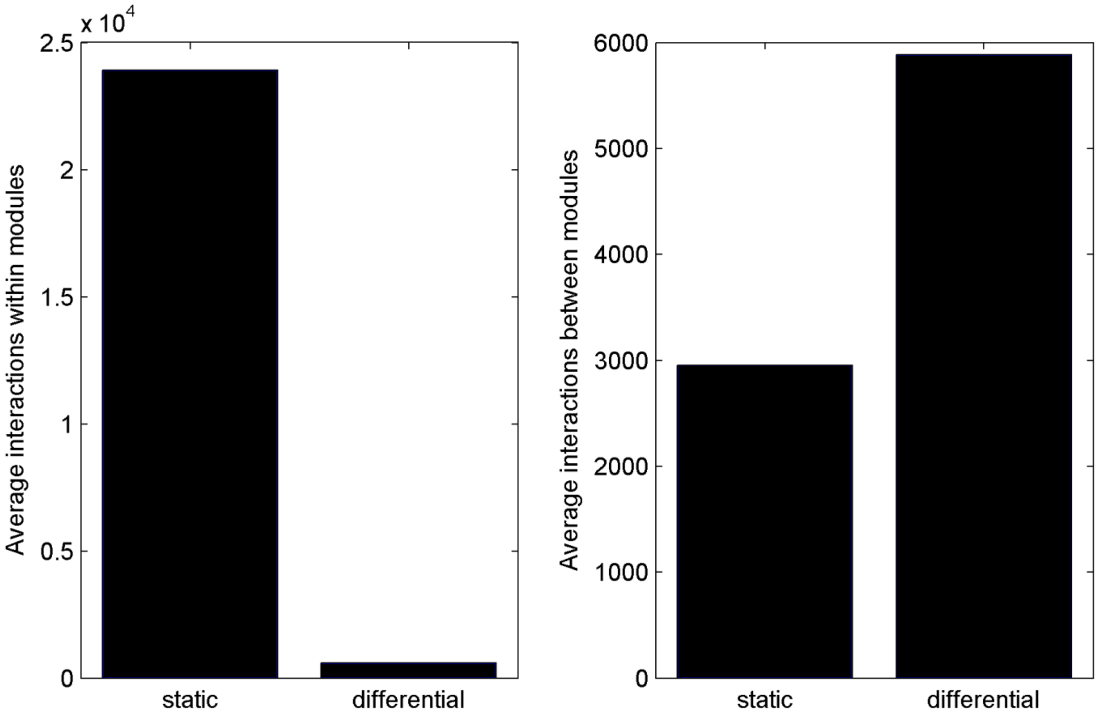


FIGURE 5.21: Portrays the number of interactions between and within modules given the static and differential experimental designs. There is greater crosstalk within modules on the static design and greater crosstalk between modules on the differential design. Using the results from both should allow one to reconstruct the circadian system.

in the Appendix. In terms of simple differential gene expression using a t-test with p value of 0.001, three lists of simple differential genes were obtained and deposited in the appendix. In terms of WT versus single mutant among the interesting GO terms were: GO:0009628 response to abiotic stimulus, GO:0009416 response to light stimulus, GO:0009639 response to red or far red light. The last category inlucded Expansin-3, FAR-RED ELONGATED HYPOCOTYL 3 and FAR-RED IMPAIRED RESPONSE 1 as the majorly affected genes. In a WT versus double mutant the most affected GO categories included ncRNA metabolic process, and cellular response to hormone stimulus, response to light and abiotic stimulus. In a single versus double mutant comparison chloroplast was the major affected cellular component and the chloroplast organization scoring highest amongst the affected processes. To explore the differences in terms of differential interactions rather than differential expression between the 3 groups, the WT, *fhy3* and *fhy3far1*, a Mann-Whitney test was carried out. The significant differences were determined by setting a conservative Bonferroni corrected threshold. The pattern of associations for each group was investigated and individual networks were examined. This relied upon Pearson correlation used as the core to detect the associations. Such experimental can be further depleted using a reduced group of direct associations yet that is not the focus here. The edges are defined through the computation of p-value

	fhy3	***fhy3far***
nodes	3026	3053
edges	109616	417723
density	0.023	0.086
avg. degree	72.08	272.05
clustering coef.	0.011	0.259

Table 5.4: Topological properties of the *fhy3* and *fhy3far* datasets.

with the null hypothesis stating that there is no correlation between the genes. Edges passing the criteria for edge inclusion were mapped in static and differential networks. On the differential scale the null hypothesis is that the correlation of 2 variables in groups is the same (conforming to 1 of the 3 previously discussed criteria) versus the alternative hypothesis being that there are differences in the 2 groups. That led to a series of interesting of networks.

Two experiments were initially performed to set the stage for the downstream analysis. Initially static networks were created on core genes. Figures 5.22 A through F portrays the geometrical representation of these networks specific per condition. It is the distribution disparsity and the localization of nodes that matters. Of note are some of the overrepresentations obtained through GO after ClusterONE was used to detect modules (presented are solely the significant ones). For WT specific this includes: GO:0009617 response to bacterium. For FHY3 specific the GO includes GO:0019748 secondary metabolic process and GO:0016137 glycoside metabolic process. For wt versus double mutant these include GO:0051537 iron, 2 sulfur cluster binding and for the double mutant vs WT specific GO:0043467 regulation of generation of precursor metabolites and energy.

These core genes were of interest to the lab hence the focus on this particular dataset and this is the focus of Figure 5.23 A through C. This figure represents the coexpression of 58 transcripts in the 3 conditions. There is resemblance between the wild type and the double mutant which led to the inspection of these specific connections. To a surprise the restablished connection are made of different nodes and edges. The fate of the nodes in subgraph A is presented as red nodes and traced and marked on subgraphs B and C. The dot product to portray the behaviour of core genes is presented in Figures 5.24 A to C. Again there is greater resemblance between the WT and double mutant versus WT and single mutant and double mutant and single mutant as if there would be ragaining of the specificity between the WT and double mutants. It is the distance between the core genes that was analysed in these subgraphs. Figure 5.25 A to F are representative

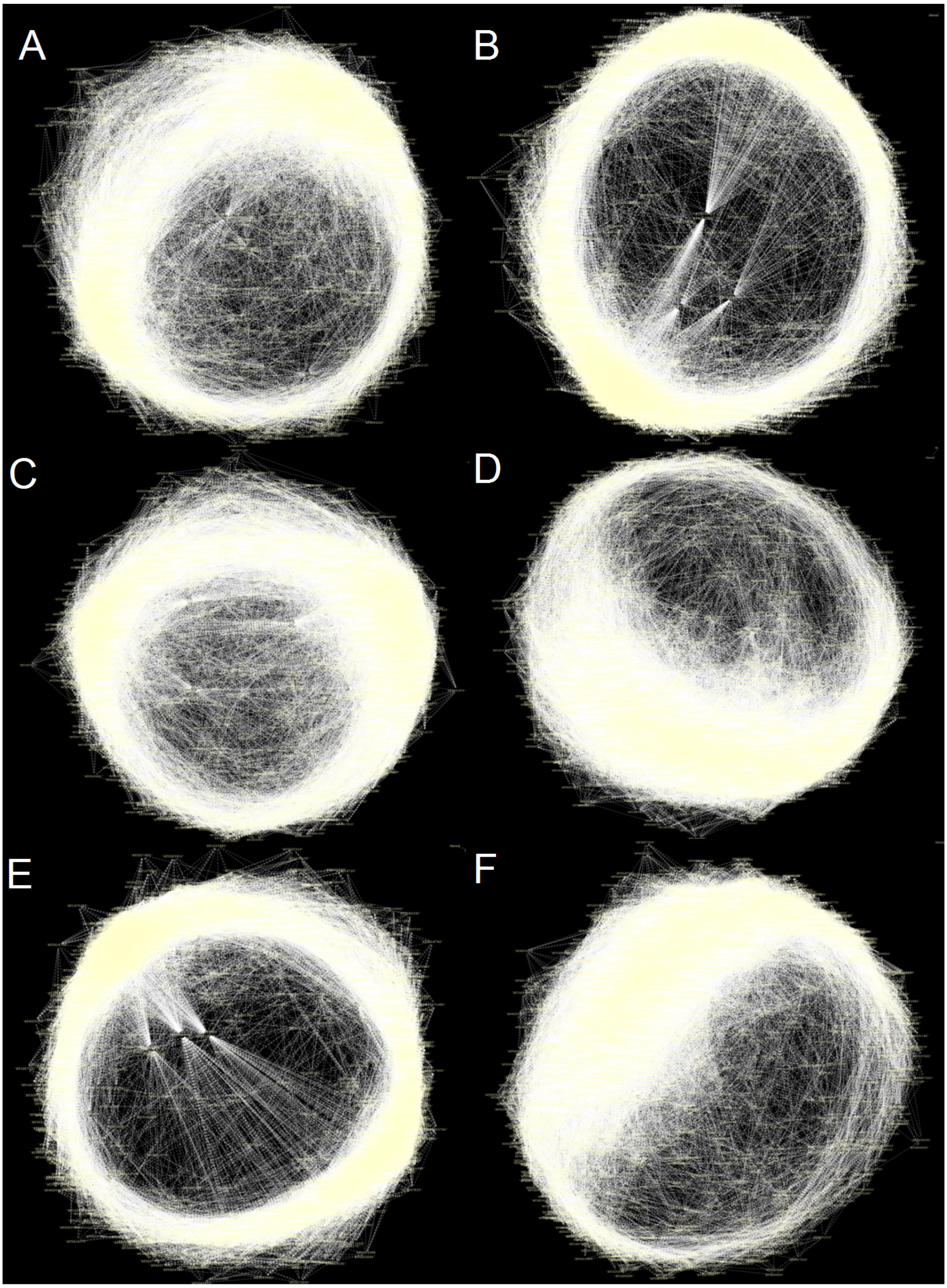


Figure 5.22: Mutant specific geometrical representation of the networks specific for the 3 conditions. In order A.WT vs *FHY3*, WT specific B.WT vs *fhy3, fhy3* specific C. WT vs *fhy3far1*, WT specific D.WT vs *fhy3far1*, *fhy3far1* specific E. *fhy3* vs *fhy3far1, fhy3* specific F. *fhy33 vs fhy3far, fhy3far* specific. It is the global topology, pattern and density that are important features of these networks, not the individual links, hence the size of the nodes.

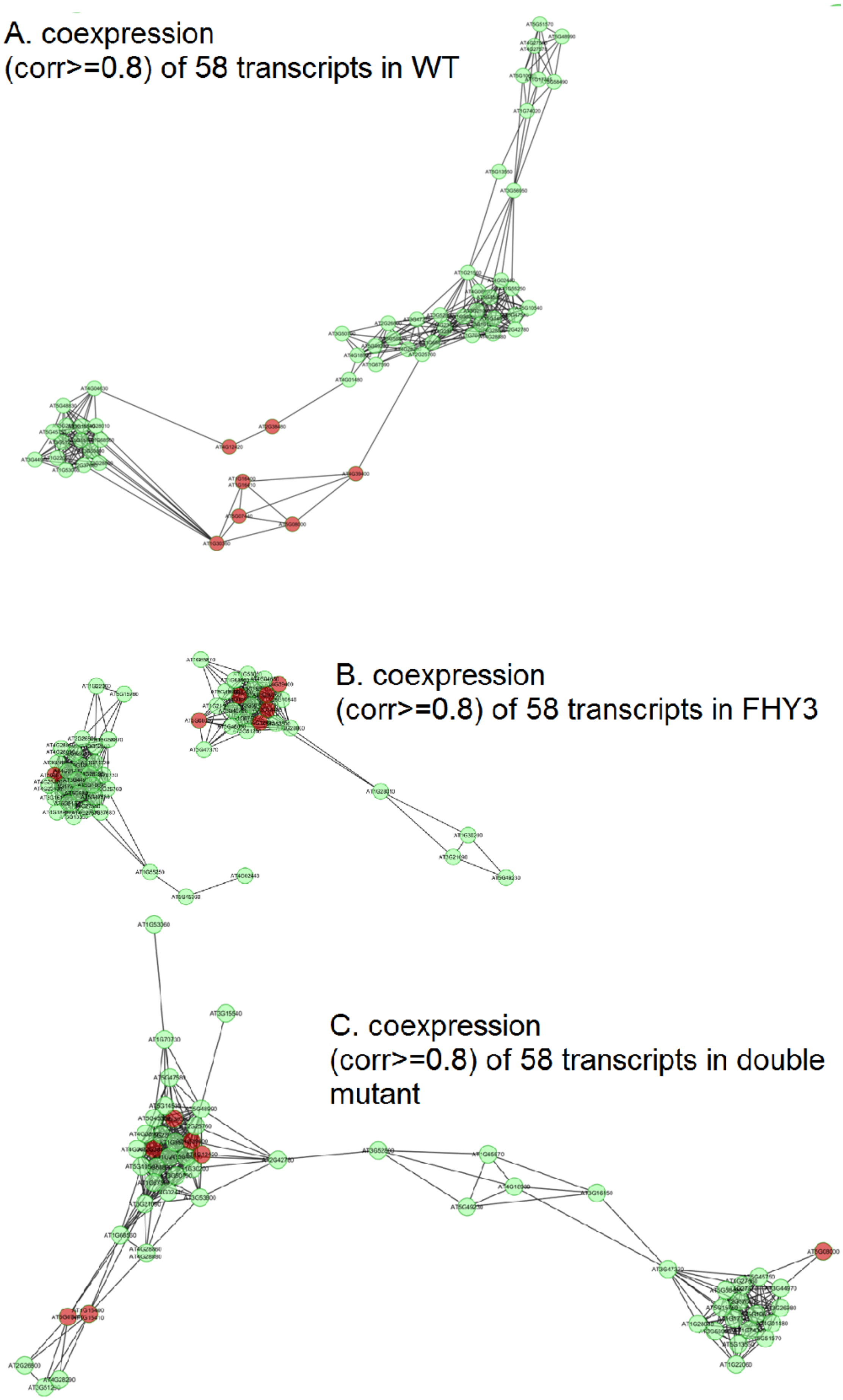

A. coexpression
(corr>=0.8) of 58 transcripts in WT
B. coexpression
(corr>=0.8) of 58 transcripts in FHY3
C. coexpression
(corr>=0.8) of 58 transcripts in double mutant

Figure 5.23: Static single and double coexpression where initially there seems to be greater similarity between the wild type and the double mutant (*fhy3far1*) versus the single mutant (*fhy3*). The nodes annotated in red are of interest as they coonect the 2 clusters in the WT condition A. That bridge is gone in panel B, single mutant and restablished in panel C, double mutant using different set of nodes.

of the initial glance at the differential topology but this time looking at the top hubs of the WT to reduce the search space. The figure demonstrates specificity for each of the 6 conditions. We initially postulated that there will be greater similarity between WT and double mutant. It is interesting to see such a connection on this group from the persepctive of CYP79F2 (P450) marked in green and atRABA1c marked in red. . Figure 5.26 portrays the static networks exemplifying major differences between the coexpression of single and double mutants. Looking at the almost identical number of nodes, the number of connections is 4 times greater in the double mutant condition, translating into a denser graph with higher connectivity. That seems to suggest the rewiring induced by the double mutant condition yet such a hypothesis would have to be verified. Figure 5.27 A to F portray on the other hand specific processes for those conditions under the selected preconditions of high correlation in 1 condition and no correlation in the other condition.

5.3 DIFFERENTIAL FINDINGS

The above allows the application of previously discussed methods to elucidation of particular motifs that represent specific groups of genes. Using the previously described methods and novel gene sets same extraction of *cis* elements was carried out. This mapping could be useful to induce changes of particular subprograms. We identified communities which constitute cohesive groups of actors that are strongly connected to each other, identified the central actors linked to many others or that bridge communities together, analysed their roles and positions with regards to different perturbations including time. We found using several datasets with alternative designs, substantial changes in interaction patterns and demonstrated that the difference in scores was more effective than the scores in either static study for highlighting interactions relevant to the pathway under study (significance testing results). Previously, it has been shown that the correlation of static interaction profiles identifies many gene functional relationships not identified by direct genetic interactions (a genetic interaction profile is the set of all interactions with a given gene). Given the quantitative score for differential interactions, we therefore investigated whether differential interaction profiles could also be used to provide distinct functional information in the circadian system. Indeed, we found that the correlation of differential interaction profiles was able to identify relationships relevant to the tested condition/phenotype/time and furthermore, that these edges were not identified either by direct interactions nor by correlation of static profiles. Additionally, one of the key limitations of static profile similarity is that the static profile is populated by interactions pertaining to both the phase as well as general conditions. Essentially

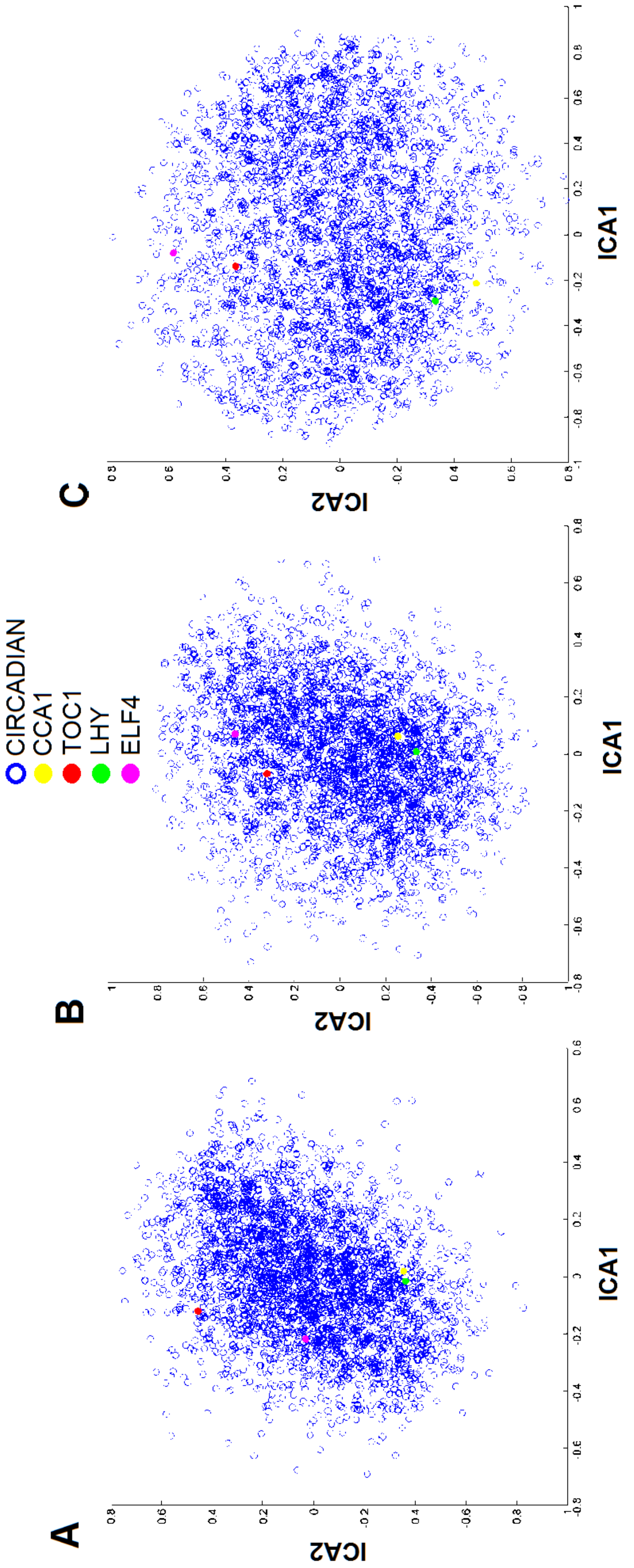


FIGURE 5.24: The core clock genes are mapped (via dot product) across the 3 conditions, wild type, single and double mutant. The axes are the top 2 components derived from the mutant microarray dataset.

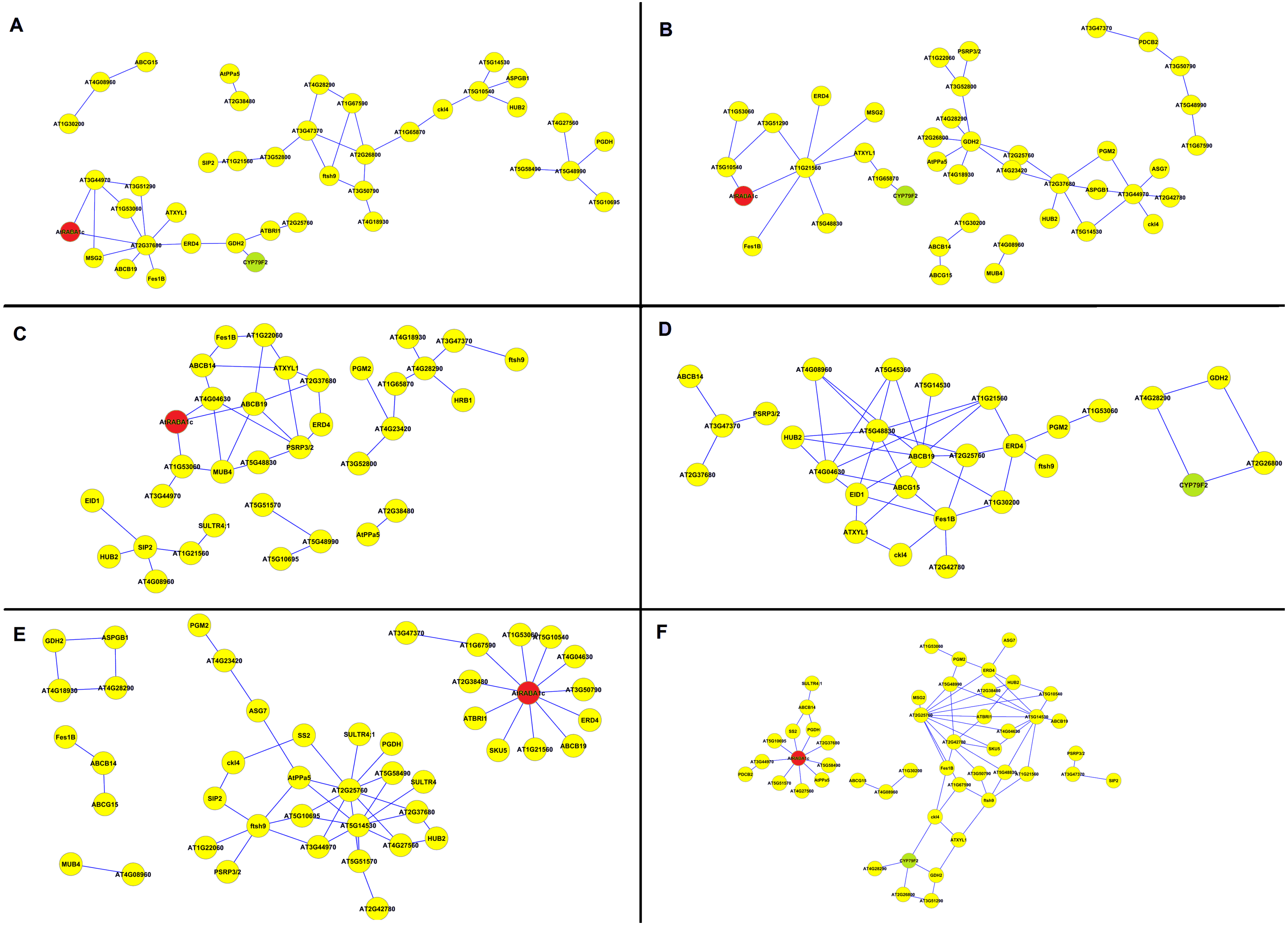


FIGURE 5.25: Differential network on 58 hubs on the mutant dataset. A.WT specific ($WT >= 0.8 fhy3 <= -0.8$) B. *fhy3* specific ($fhy3 >= 0.8 WT <= -0.8$) C. WT specific ($WT >= 0.8 fhy3far1 <= -0.8$) D. *fhy3far1* specific ($fhy3far1 >= 0.8 WT <= -0.8$) E. *fhy33* specific ($fhy3 >= 0.8 fhy3far1 <= -0.8$) F. *fhy3far1* specific ($fhy33far1 >= 0.8 fhy3 <= -0.8$) CYP79F2 (P450) is marked in green and atRABA1c is marked in red.

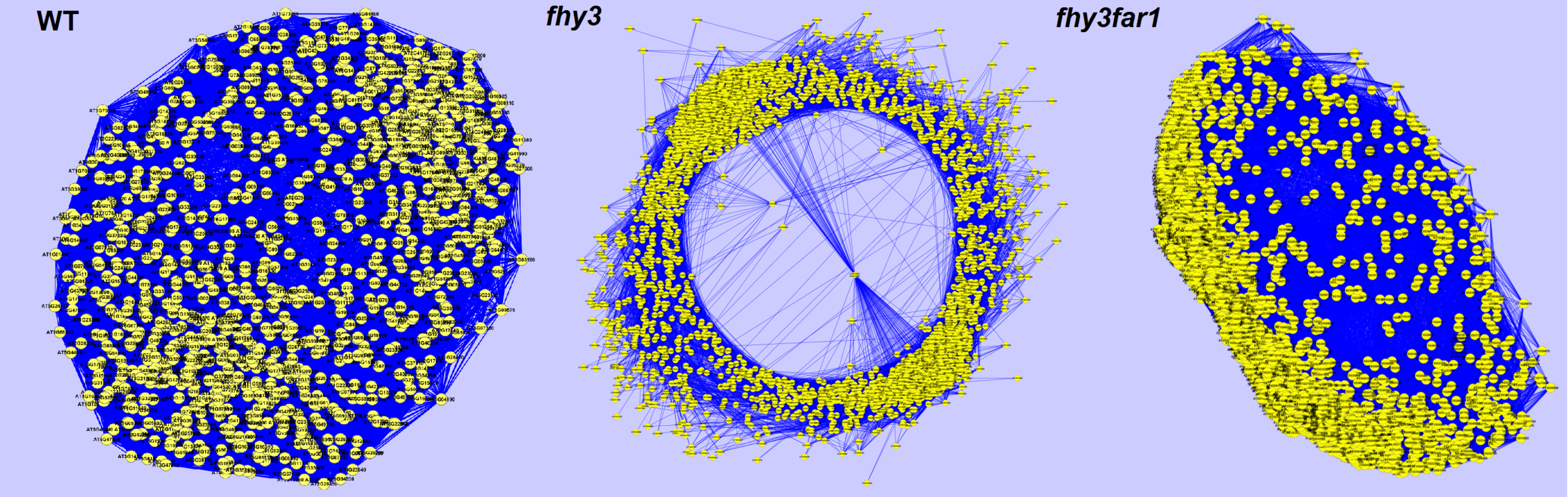


FIGURE 5.26: Individual static networks for WT, single and double mutants where each of the connections indicates a significant pairwise association between the genes using a corrected threshold $p < 0.01/z$, where $z = M(M-1)/2$ and M is the number of genes. It is the global topology, pattern and density that are important features of these networks, not the individual links, hence the size of the nodes.

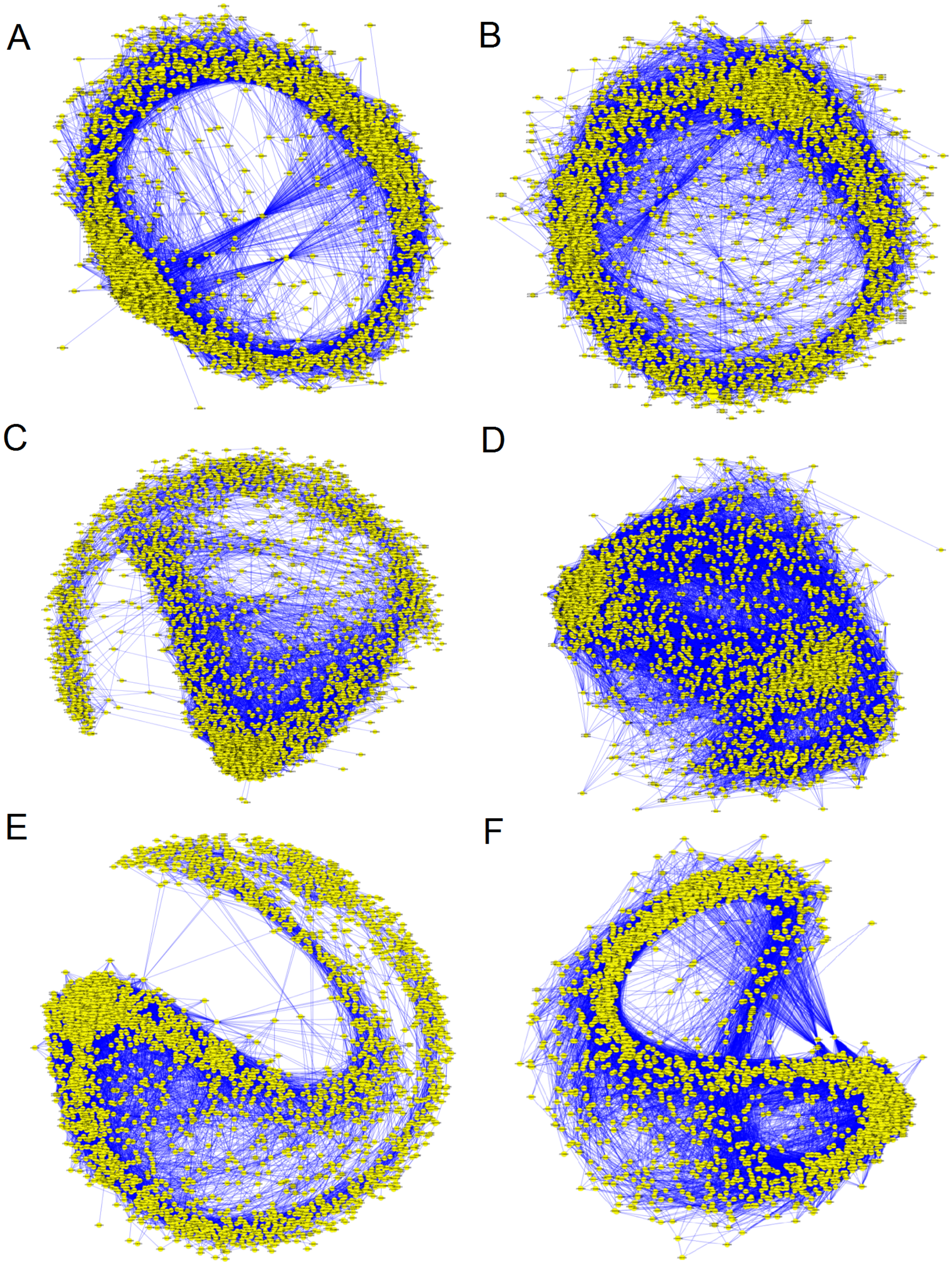


FIGURE 5.27: Mutant dataset portrayed in terms of differential interaction as placed upon a statistical footing. All the networks were prepared in Cytoscape using the same edge weighted spring embedded layout. The selection criteria are defined by method 3 (08 (02.02)). A. *fhy3* vs WT, *fhy3* specific B. *fhy3* vs WT, WT specific C. double vs WT, double specific D. double vs WT, WT specific, E. single vs double mutant, double mutant specific F. single vs double mutant, single mutant specific. It is the global topology, pattern and density that are important features of these networks, not the individual links, hence the size of the nodes. Individual connections are discussed in the text

the larger variance inherent in the static measurements contributes to noisier interaction profiles which decreases the similarity of otherwise related profiles. Differential interactions are strong and effective at identifying time and mutant relevant relationships because they cut down the noise and eliminate unrelated interactions. With the hypothesis being that particular hubs will drive programs at particular times yet not others, here the hubs at particular times can be seen as the hubs. One such are TF; these can be used for identifying the core time programmes. Molecular profiles are complex and information-rich, calling for cost effective tools to analyse, visualize and compare the underlying biological processes. A formal statistical approach for the differential analysis of associations via network representation has been used with a demonstration of its biological application to assess changes in molecular associations between different conditions. The approach introduced here can provide insight into the biological basis of phenotypic variation and aid the generation of new hypotheses about the circadian control and regulation in the context of systems biology.

Altogether, the modules identified using eigenanalysis taken together with the central hubs and clusters within these differential programs can remap the system from a novel angle, the sole experiment is essentially validation at a greater resolution to determine more specific *in silico* rearrangements. Time will show how effective this will be in identifying and next changing these modular programs.

This chapter shows how a development of methodology can serve great benefits for a circadian system. Selection 1-2-1 for example can delineate the interactions specific to WT and a double mutant not present in a single mutant. Furthermore such novel interation design allows for one more powerful *cis* element grouping. It revealed a highly interesting combinaton of the EE like element with GATA element across different phases. From the perspective of *cis* regulatory modules, that has direct implications, these being ability to determine motifs characteristic of cliques. Here these findings will further serve in Chapter 6 where the focus is on TFs per se. Nevertheless the data would greatly benefit from greater resolution as then the same principle could be applied to phases. It is challenging to decode network dynamics from the multi-conditional gene expression data due to reasons like the diverse activation patterns of nodes over multiple conditions, the dense interactions (edges) among the nodes; and the large size of the global network. All cellular functions and fate decisions are governed by spatiotemporal design principles: circadian rhythms, cell division, development and metabolism. There is a notion that the temporal complexity scales with structural complexity: more complex organisms call for increased regulation of various biological processes, hence the emergence of different levels of time-coupled responses. Modularity is clearly seen in periodic processes other than the circadian clock like the cell cycle. Through studying the dynamic patterns of biological processes we identified the control points, modules and sources of robustness. Here the circular depictions were translated into layers that enable simultaneous comparisons of network states at different time points. We anticipate that these aspects will be better integrated in the future by using a temporal framework that combines different modelling techniques, These results lay the foundations for a large-scale approach of predicting phenotypes based on the schematic structure of networks. This had already been attempted when the effect of 39 perturbations was predicted upon 433 target proteins/genes. In up to 82% of the cases, an algorithm that used only the static structure of the network correctly predicted whether any given protein/gene is upregulated or downregulated as a result of perturbations of other proteins/genes (Feiglin et al., 2012). Interactomes from BioGRID and BIND do have a large element of false positives and false negatives nevertheless the majority constitutes important information (Bader et al., 2003, Stark et al., 2011).

5.4 DISCUSSION

First and foremost the differential interactions uncover many gene functions that go undetected in static conditions. One direct example is looking at the top results of a differential expression analysis between wild type and *fhy3* versus the differential interactions example for that specific combination. The same holds true for the double mutant *fhy3far1* under red light. At3g57930 (a hypothetical protein for the time being) was the top DE protein in both settings. In terms of the *fhy3* versus wildtype set, among to top differentially expressed genes were *LHY*, *COL2* and *RVE8*, *HYH* and *CCA1* with overrepresented GO terms being response to light stimulus and KEGG pathway being circadian rhythm. For the double mutant among the differentially expressed genes were *ELF4*, *CCA1*, *GRP9* and B12D protein and *LHY*. Regarding the processes these include plant cell wall organization ranked highest in terms of enrichment and photoperiodism, flowering. It is not that one is better than the other, it is a different perspective upon regulation particularly interesting from the juxtaposition of time series. It is worth mentioning that *FHY3* has beem associated with flowering and photosynthesis and even with branching but never with cell walls. Likely it was shown that protein complexes are stable when perturbed but the functional relations between these are reorganized.The *fhy3* (for far-red elongated hypocotyl3) mutant of Arabidopsis *thaliana* is involved in independently gating signaling from a group of photoreceptors to an individual response (Allen et al., 2006). Alone the *fhy3* mutant phenotype is manifested by short hypocotyl and loss of circadian rhythms under red light. It was proposed that *FHY3* and *FAR1* in concert represent TF that have been co-opted from an ancient Mutator-like transposases to affect te phyA-signaling homeostasis in higher plants (Wang et al., 2011). One of the range of phenotypes is manifested by for example short hypocotyls (Lin et al., 2007). Here, through network subtraction, the housekeeping interactions are depleted allowing for detection of the pathways that are differentially represented. One direct finding is that whereas static genetic interactions are enriched within modules, we found that differential genetic interactions are not. Instead those differential genetic interactions are much more likely to occur among pairs of genes connecting two different modules than among pairs of genes within the same module. Those modules which are the dense clusters of static interaction were the initial building blocks. The module module interactions were characterized by greater enrichment for many differential genetic interactions across the two modules. This is incredibly important in terms of a day and night network as clearly we are aware that these are going to be different yet now we know how so. Essentially, a different set of interactions governing a dynamic cellular response was captured. Given that most gene functions arise in response to changing conditions, the differential network revealed here offers a interesting view upon particularly pertaining to time snapshots. The answer is yes indeed to all 3 questions that were evaluated with high confidence. These are whether the overall modular structures of the two networks are different; whether the connectivity of a particular set of "interesting genes" has changed between the two networks, and whether the connectivity of a given single gene has changed between the two networks. Using the same network layouts as seen throughout the differential plots and through elimination of the static interactions a totally novel gene hubs were inferred like the drivers of the night and the day. Such control and hub nodes and their connections lead to smaller *cis* groups of regulation. Knowing their regulatory targets allows combining this information to alter the time expression using the uncovered *cis* element. It would be incredibly fascinating to work a high resolution dataset which would allow one to apply such differential analysis to the 4 modules, to several phases. Network concepts are important indices for understanding complex systems. For example, the clustering coefficient has been used to understand network architecture; the connectivity has been used to screen for cancer targets whereas and the topological overlap matrix has been used to define modules and to annotate genes. In the present chapter, we looked at functional enrichment at the structure of cellular networks with the aim being ability to model regulatory networks to study the dynamic behaviour of gene regulatory networks and to alter it. Connectivity is important as the hub genes like for example *RVE7*, *PIL6*, *TCP11* and *MPK7* are thought to play an important role in organizing the behaviour of differential day and night biological networks. The present hubs for static networks include *NPX1* for the day and *RVE8* for the night. The centralization index has been used to describe structural differences between networks. Biological networks tend to be very heterogeneous, whereas some 'hub' nodes are highly connected, the majority of nodes tend to have very few connections. Modules were shown to be approximately factorizable, have high density and high conformity, high centralization and high average clustering coefficient as shown by (Dong and Horvath, 2007) (these properties are another index of their detection). Functional annotation useful in itself was used as a validation of the subnetworks. A deeper understanding of network concepts should guide future alteration of such networks. All those indices explored in the present chapter affect the understanding of the clock network. There are geometric reasons why highly connected hubs like *RVE8* for example in important coexpression modules tend to be important, and why hub genes in one module cannot be hubs in another distinct module. Such a module centric analysis alleviates the multiple testing problem deeply embedded in microarray data analysis. Instead of relating thousands of genes to a sample trait, it focuses on the relationship between several such modules and the sample trait. In terms of the geometric interpretation of the hub gene significance the smaller the angle between the module eigengene and the clock structure, the higher is the hub gene significance. In principle, the genes being intermediate between two distinct modules cannot be a highly connected intramodular hub gene in either module. Essentially, the large angle between module eigengenes reflects that the corresponding modules are distinct. Since intermediate genes do not have a small angles with either eigengene, thye cannot be classified as intramodular hub genes. That principle was the driving factor for the static analysis which uncovered the potential novel candidates for altering the phase outputs.

As these methods develop linking multiple types of omic data will be instrumental to the understanding of the system. It is the study of community structure of multislice networks, these being combinations of networks coupled through additional links that will provide the true dynamics of the system. That should allow one to comprehensively study networks that evolve in time as proposed by Mucha (Mucha et al., 2010). Instead of studying one static network at a time and then attempting to assemble the pieces together, the overlay method allows one to study the network community structure in terms of quality functions using multiple types of links simultaneously. Classically the method for quantifying the notion of community is through the use of a quality function, counting the intra versus inter community edges and comparison to a random model. This has been limited to a single static adjacency matrix. As different spreading weights are allowed on different types of links the formalisms are extended to multislice networks, a focus of chapter 6.

Chapter 6

DIALOGUE CONCERNING THE CIRCADIAN TRANSCRIPTION FACTOR TARGET REGULATORY NETWORK

6.1 INTRODUCTION

In this chapter a regulatory network based on circadian transcription factors was created. The proper working of circadian regulation is paramount to plant functioning, and is heavily reliant on the action of transcription factors. The list of circadian transcription factors was obtained based on the selected circadian genes in Chapter 3. A fundamental conundrum for regulatory networks is to decipher the relation between form and function. The objective is to uncover the underlying design principles of the gene regulatory network. Circadian clocks present a particularly interesting instance, as recent work has shown that they have complex structures involving multiple interconnected feedback loops and they operate at multiple levels of regulation. Transcription factors play a central role in the regulation of gene expression. Their interaction with specific elements in the DNA mediates dynamic changes in transcriptional activity. These are central to the understanding of the plant clock and hence particular attention is placed on a series of them acting alone and with others, from the perspective of families and modules. This provides a comprehensive picture of the regulatory and mechanistic pathways contributing to circadian function. This chapter builds upon the findings of chapters 3,4 and 5 yet the analysis is from a different angle, that is, emphasis is placed on the individual transcription factors and their cliques. The transition is from co-expression networks that are largely association networks to gene regulatory networks (GRN) where directed edges might translate to causal relationships. Transcriptional regulatory networks describe how transcription factors control the expression of target genes. The nodes are the TFs and the edges illustrate the regulation between them. Network theory has been successfully employed to study the genetic coexpression in plants allowing for predictions and the discovery of functionally related genes from the underlying data. Throughout the analysis both promoter binding and activation via TFs is important, yet at times it is simply the presence of a particular agent that is critical for the functioning of the complex. It has to be kept in mind that these are time series data being exploited. In these models of gene regulatory networks, oscillations originate from a combination of negative feedback mechanisms and time delays. That is, for a single-gene, autorepression loop does not oscillate unless either a time delay is introduced versus multiple enzymatic degradations are introduced resulting in high nonlinearities. Nevertheless, many models were geared towards capturing linear relationships. A system with three and more components, like the Goodwin oscillator, is able to oscillate as a result of sufficiently high nonlinearity combined with a large number of components at play (Woller et al., 2013). Uncovering the intracies of these loops calls for a set of assumptions like those regarding the time lag. The details of the GRN inference methods used to infer the TF targets, and the models created will be discussed in depth.

6.2 TOOLS FOR GENE REGULATORY NETWORK INFERENCE

Modelling of the gene regulatory networks goes back to 1960s and the pioneering work of Stuart Kauffman and Rene Thomas (Gjuvsland et al., 2007). 'Reverse-engineering' can be defined as the process of identifying gene interactions from experimental data through computational analysis. Gene expression data from microarrays are typically used for this purpose yet that is now being replaced or supplemented by NGS technologies. The aim is to select a method that will tackle the problem at hand. Several different reverse-engineering algorithms have been tested on experimental data sets. What is rewarding is that reverse-engineering algorithms are indeed able to correctly infer regulatory interactions among genes. With the avalanche of data comes a selection of methods and algorithms that were developed to define the key connections. That is best portrayed by the DREAM (Dialogue for Reverse Engineering Assessments and Methods) initiative (Stolovitzky et al., 2007). It aims to evaluate the performance of GRN inference algorithms on benchmarks of simulated data serving as excellent confirmatory tool. This deluge of information can be best represented as the transcriptional regulatory network with nodes connected by edges. In such a network representation, the nodes represent either the transcription factors or their targets where the directed edges represent a regulatory interaction (protein and DNA interaction) between the TFs and their targets (TGs). Depending on the promoter strength of the regulated genes, it may respond to different concentration levels of the active TF which is the key point here. Therefore, if the concentration of the active TF changes with time, such a motif could set a temporal pattern in the expression of the individual targets. Globally, the analysis of transcriptional networks has revealed that they display a scale-free topology; that is, they are characterized by the presence of a few highly influential TFs that regulate many genes and a large number of TFs that regulate only a few genes. The highly influential TFs are referred to as global regulators, the hubs, and their presence contributes to the inherent robustness of such a topology. Here robustness is defined as the ability of complex systems to function even when the structure of the system is perturbed significantly. As described in Faith and Gardner et al., (2005), there are two broad classes of reverse-engineering algorithms (Gardner et al., 2000, Gardner and Faith, 2005). The first rely on the 'physical interaction' approach that aim at identifying interactions among transcription factors and their target genes (gene-to-sequence interaction) and the second are those based on the 'influence interaction' approach that try to relate the expression of a gene to the expression of the other genes in the cell, gene to gene interaction, rather than relating it to sequence motifs found in its promoter, gene to sequence. Nevertheless, the presence of interaction between two genes in a gene network does not necessarily imply a physical interaction, as it can refer to an indirect regulation via proteins, metabolites and ncRNA that have not been measured directly. In general, however, the meaning of influence interactions is not well defined and depends on the mathematical formalism used to model the network. Nonetheless, influence networks do have practical utility for identifying functional modules; that is, identify the subset of genes that regulate each other with multiple (indirect) interactions, but have few connections to other genes outside the subset. Gene network models can be used to predict the response of a network to an external perturbation and to identify the genes directly disturbed. Additionally, identifying real physical interactions by integrating the gene network with additional information from sequence data and other experimental data like chromatin immunoprecipitation can be done.

Understanding gene regulatory networks provides in depth understanding of the working of system at the molecular level allowing for identification of potential targets for genetic modifications. To overcome the challenges associated with this inference, a number of competing approaches have previously been used, including those from information theory, Bayesian and Dynamic Bayesian Networks (DBNs), Ordinary Differential Equations (ODE) and stochastic differential equations, regression based methods to name a few (cite). The reconstruction of regulatory networks for the circadian purpose of this chapter involved several trial experiements that included the implementation of ARACNE (Margolin et al., 2006), timedelayARACNE (Zoppoli et al., 2010), IOTA (Hempel et al., 2011) and Granger (Krishna et al., 2010) causality testing. The focus of these sections is on key methods and their potential applications. In plants the statistical methods for the construction of gene coexpression regulatory networks (ie TF:TG networks) are mainly focused on the Pearson Correlation Coefficient (PCC) approach, yet these have also been supplemented by Graphical Gaussian Models, Bayesian Networks, Dynamic Bayesion Networks and information theoretic approaches.

Pearson correlation is the most widely used inference methodology in plants calculated based on the expression values of genes either across many experiments versus across time. The metric scores the tendency of two genes to show similar expression levels across samples. The expression levels from pairs of genes with a larger correlation value than a threshold are considered to reveal a potential interaction, influence, dependence or coordinated participation in the same function (López-Kleine et al., 2013). For example, a study performed by Mao examined all the modules within 1094 ATH1 arrays using PCC with the aim of comparing these findings and inferring novel biological roles. The simplicity of this measure and the ease of use has tempted many researchers yet that comes at a tradeoff of loss of information for example due to nonlinear interactions. The graphical Gaussian model (GGM) is an alternative method to the Pearson correlation. GGM-based methods are undirected probabilistic graphical models which describe the conditional independence relationship among genes under the assumption of a multivariate Gaussian distribution of the data. In the GGM networks, each node represents a gene and an edge connects two genes if they are partially correlated. Essentially, the GGM calculates the empirical covariance matrix from a dataset that is then inverted, after which the partial correlations are computed. It does possess several strong advantages as for example it considers the effect of other genes upon the connection of interest. The fundamental disadvantage is that it is best suited for cases in which the number of samples N is relatively large compared with the number of variables p. This is not the case here where N is by far smaller than p resulting in the correlation matrix not having a full rank and not being able to be inverted (has a determinant of 0). Nevertheless, there is a interesting plant example where such a model was used with the focus on starch metabolism in leaves (Ingkasuwan et al., 2012). The assignment of genes relied upon their day and night expression pattern and was carried out to identify particular patterns of co-regulation with starch biosynthesis and degradation, indicating a relationship between TFs targeting starch metabolic genes (TFs and their TGs). There are many alternative inference methods of which one is the construction reliant upon the Bayesian theory to represent the probabilistic relationships between all genes as initially proposed by (Friedman, 2000). Bayesian network constitutes a graph-based model representation of joint multivariate probability distribution. It captures properties of conditional independence between variables. This leads to static or dynamic Bayesian networks (BN). In biological terms, Bayesian network decomposes the joint probability of the expression values of genes in the network into the probability of the expression value of each gene, given its regulators. Indirect relationships can be captured. BN methods are more resistant to noise than for example Boolean networks. In Boolean networks, logical function is written as a statement acting on the inputs using the logical operators: and, or and not (Albert, 2004). Its output is 1 or 0 if the statement is true or false. The problem is that transcription is not a on/off process. A nice portrayal of the applicability of the dynamic BN was shown in the work done by Dondelinger et al., where nine circadian genes were used to produce a model of gene regulation within the plant circadian clock (Dondelinger et al., 2012). Limitations exist for the classic BN, as for example the fact that, finding causal relationships may be impeded by the fact that certain networks are equivalent yet exactly these causal relationships can be detected in the dynamic BN. In this study, individual network was developed for four circadian microarray experiments, produced under varying conditions and a novel information sharing method was applied to merge the networks. As a result it has been shown that it is indeed possible to reconstruct known gene interactions with the right assumptions of the applied methodology. Unfortunately, such inference is an NP-hard problem, and BNs perform best with small networks from tens to hundred nodes.

Information theoretic approaches constitute another domain of possibilities for inference. In principle, the regulatory interaction between two genes is established if the mutual information on their expression patterns is significantly greater than a P- value calculated from the mutual information (MI) between random permutations of the same patterns. The classic example is a relevance network. Essentially the information theoretic approaches use MI, to compare expression profiles from a set of microarrays. For each pair of genes, their MI_{ij} is computed and the edge $a_{ij} = a_{ji}$ is set to 0 or 1 depending on a significance threshold to which MI_{ij} is compared. MI can be used to measure the degree of independence between two genes. MI becomes zero if the two variables x_i and x_j are statistically independent. A higher MI indicates that the two genes are non-randomly associated to each other. MI captures nonlinear interactions which is advantageous over simple correlation. Edges in networks derived by information-theoretic approaches represent statistical dependences among gene expression profiles. As in the case of Bayesian network, the edge does not represent a direct causal interaction between two genes, but only a statistical dependency. Algorithm for the 'Reconstruction of Accurate Cellular Networks' known as ARACNE and its extension known as TimedelayARACNE consitute tools that operate using MI. ARACNE is a relevance network approach, yet it introduces additional scoring rules for the pairwise weighting of the interactions with the aim of reducing the amount of links that are falsely detected (Margolin et al., 2006, Zoppoli et al., 2010). It computes M_{ij} for all pairs of genes i and j in the data set. M_{ij} is estimated using the method of Gaussian kernel density. Once M_{ij} for all gene pairs has been computed, ARACNE excludes all the pairs for which the null hypothesis of mutually independent genes cannot be ruled out ($Ho : MI_{ij} = 0$). A P-value for the null hypothesis, computed using Monte Carlo simulations, is associated to each mutual information value. The final step of this algorithm is a pruning step that tries to reduce the number of false-positives. The method applied the data processing inequality (DPI) principle that asserts that if both (i, j) and (j, k) are directly interacting, and (i, k) is indirectly interacting through j, then M_i, $kmin(M_{ij}, M_{jk})$. It is not suited though for short time series nor is it useful for time series, hence timedelayARACNE was developed as the extension of the model (Zoppoli et al., 2010). TimedelayARACNE is a dynamic information theoretic approach that uses MI to infer interactions. It may be beneficial for small networks yet has limitations in terms of larger networks like that circadian of 300 TF and 3070 potential targets.

The Granger causality testing was yet another powerful approach having many advantages like foremost addressing the directionality of couplings yet, even after focusing on previously annotated interactions, yielded networks with over 10,000 edges which is too dense to be of use (Krishna et al., 2010). The settings were set to identify the current regulators; that is, for the present timepoint and the two preceding ones relying on verified findings and distribution analysis. These were taken into account when fitting the autoregressive model in the Granger causality (GCT) initially applied in a different domain of the sciences (Granger, 1969). In a grand study comparing multiple methods Granger causality led to close to random prediction of links. Hence, the Granger causality measure became not suitable for the reconstruction of such a GRN, particularly considering the short gene expression time series are available. The cause is that the results of the GC index rely heavily on the model estimation (Hempel et al., 2011). Given we have 12 time points over 48 hours, then each time point corresponds to 4 hours, and using the autoregressive model of order 2 corresponds to 8 hours, meaning that you consider a causal relation spanning 8 hours as significant (TF and targets). ClusterONE was used to rank and discriminate between the subnetworks (Nepusz et al., 2012). Cluster ONE filters away clusters with a density of the network less than 0.2 by default, but since the weights obtained from GCT are large, even a single edge with a very high weight can easily push the density of a cluster over the threshold. Therefore logistic transformation was used. That consisted of calculating the median of the weights, using logistic mapping that maps the median weight to 0.5. This was applied to edges with non-zero weight, which ensures that the sparsity remains. The results were dense and the methodology was inconclusive in this particular design hence alternatives were considered.

In the present analysis, another measure of dependence for two-variable relationships was tested called the maximal information coefficient (MIC) (Reshef et al., 2011). It captures a wide range of associations, these being both functional and not (based on the tested data). For functional relationships (links pertain to the same functionality), provides a score similar to the coefficient of determination (R^2). The measure should capture a wide range of other interesting associations including, yet not limited to, function types (linear, exponential, periodic). It performs best with detection of non-linear relationships between variables with moderate amounts of noise. It is, furthermore, a highly equitable statistic, meaning that it should output similar score to equally noisy relationships of different types. There are many methods that detect a range of relationships including simple mutual information, Spearman rank correlation coefficient and principal curve-based methods to name a few. Nevertheless, they are not equitable, showing greater preference for one type of function. It was successfully applied to the reconstruction of a T cell signalling network from single cell data as shown by (Yosef, 2013). In this chapter, the measure was used and compared to simple mutual information in reconstructing the regulatory network.

Ensemble methods constitue another type of GRN inference strategy. The main idea is to consists of several steps: bootstrap a given data set, apply a network inference method, and aggregate all separate outcomes into a final result (Huynh-Thu et al., 2010). Tree based regression methods constitute a working example of GRN inference. One such particularly interesting method and winner of one of the DREAM challanges was GENIE3 (Huynh-Thu et al., 2010). The beauty of Tree based ensemble methods lies in the fact that they do not make any assumptions about the underlying nature of the target function. Using GENIE3, RF is trained to predict target gene expression. It operates on the principle that TFs are selected as tree nodes if they consistently reduce the variance of the target gene. Per each gene, a learning sample is generated with expression levels of j as output values and expression levels of remaining genes as input values (Huynh-Thu et al., 2010). A local ranking of of all genes minus one is computed. These are then aggregated to get a global ranking of links. The method due to its effectiveness in recovering true causal links was one of the methods of choice and will be presented in the results section.

Synthetic data enables one to inspect the performance of algorithms against a ground truth. Moreover, when many of these methods were compared on the same data, the different reverse-engineering methods considered here inferred networks that overlapped for about 10% of the edges for small networks, and even less for larger networks (Marbach et al., 2010). Interestingly, however, if all algorithms agree on an interaction between two genes (an edge in the network), this interaction is not more likely to be true than the ones inferred by a single algorithm. Therefore, it is not a good idea to consider a interaction as true positve simply because more than one reverse-engineering algorithm recovered it. Initiatives like the Dialogue on Reverse Engineering Assessment allow for better assessment of the multiple tools yet it has been shown that no single inference method performs optimally across all data sets (Marbach et al., 2010). In a comprehensive study performed by Kurt in 2014, altogether 27 correlation-based and MI-based interaction estimators were evaluated, with the conclusion that even though some perform better there is no one best suitable method, so the choice depends totally on the problem at hand. The top estimators altogether include B-spline, Pearson correlation and Spearman based methods according to the study. If there is no correct method, the choice has to be made depending on the type of the data, size of the network, number of samples, level of noise, the experimental design, type of network structure (in example scale-free), the error measure and likely other features.

In this chapter we set out to infer a TF:TG network for the plant clock. Several of aforementioned methods were tested. We rejected pure PCC on the basis that in PCC the similarity between genes relies only on linear relationships, whereas nonlinear relationships remain undetected. The method is sensitive to outliers, generates many false positives and does not distinguish indirect and direct relationships. On the other hand, GGM eliminates these indirect effects yet this was also rejected as it does so at a computational cost (the circadian network is too big). In the case of BN, the same problem of computational cost applies hence the method of choice was a information theoretic approach and the ensemble apporach presented by GENIE3.

6.3 RESULTS: DIFFERENT INTERACTION ESTIMATORS AND SCORING SCHEMES

We address the problem of recovering regulatory network from circadian gene expression data. Specifically the aim was to infer a transcription factor: target regulatory network, using a information theoretic approach reliant upon mutual information and chosen based on extensive testing. The resulting network and its applicabilitiy will be discussed in the subsections to follow. Throughout the chapter, the following notations will be referred to: the gene expression measurement of gene i with the variable x_i, the set of expression measurements for all the genes with D and the interaction between genes i and j with a_{ij}(details in Chapter 2). D consists of time-series gene expression data of N genes in M time points, with gene expression changing dynamically with time. Usually the gene network can be either an undirected graph, that is, the direction of the interaction is not specified ($a_{ij} = a_{ji}$), or a directed graph specifying the direction of the interaction, that is, gene j regulates gene i (and not vice versa) ($a_{ij}a_{ji}$). The design involves 300 TF derived from the list of circadian genes using the selection process described in chapter 3. The direction is specified as we are looking for targets of those 300 circadian TF. The list of the TF is provided in the appendix and the distribution with respect to time of peak expression of certain TF families, those prevalent to the clock are seen in Figure 6.1. The figure portrays the location of TF on a component dot product plot belonging to the 4 families directly relevant to the circadian clock. Whereas there is bimodal distribution of the bHLH transcription factors ad the TCPs, there is a half cycle continuous distribution of the MYB and the MYB related transcription factors (spanning across different times of the day, with greater density during dawn and dusk). The day and night would explain the striking bimodal distribution seen within two of the families. It is clear that the distributions will vary accordingly with functional groups yet what is of interest are the dependcies between the outputs of these TFs.

A directed graph can be labelled with a sign and strength for each interaction, signed directed graph, where a_{ij} has a positive, zero or negative value indicating activation, no interaction and repression. In order the select the optimal method for the circadian TF regulatory network, many were tested with prime examples being the Granger causality, ARACNE and timedelayARACNE and modified information theoretic approaches as discussed in the introduction. The justification for not using a purely linear approach is presented in Figure 6.2 where some existing known interactions are portrayed as scatter plots of expressions of these two genes. Furthermore Figure 6.2 exemplifies the applicability of time delay methods in capturing the interactions between TF and their TGs. The relationship between *COP1* and *HY5* and that of *CHE* and *CCA1* is shown on the figure. In the former, correlation captures the relationship whereas in the latter it does not, clearly due to nonlinear nature of the relationship. While "target" usually refers to the genes whose promoters are bound by a transcription factor, in this case HY5 is a transcription factor targeted by a ubiquitin ligase, COP1. This does not mean "target" in the usual sense. It was shown that COP1 negatively regulates HY5 which is a positive regulator of photomorphogenic development. It interacts directly and specifically with HY5 (in vivo and in vitro assays) (Biogrid inferred data). Here mutual information and MIC are able to capture diverse types of relationships. The use of both is supported in the data whereas A is clearly linear and B clearly is not (mutual

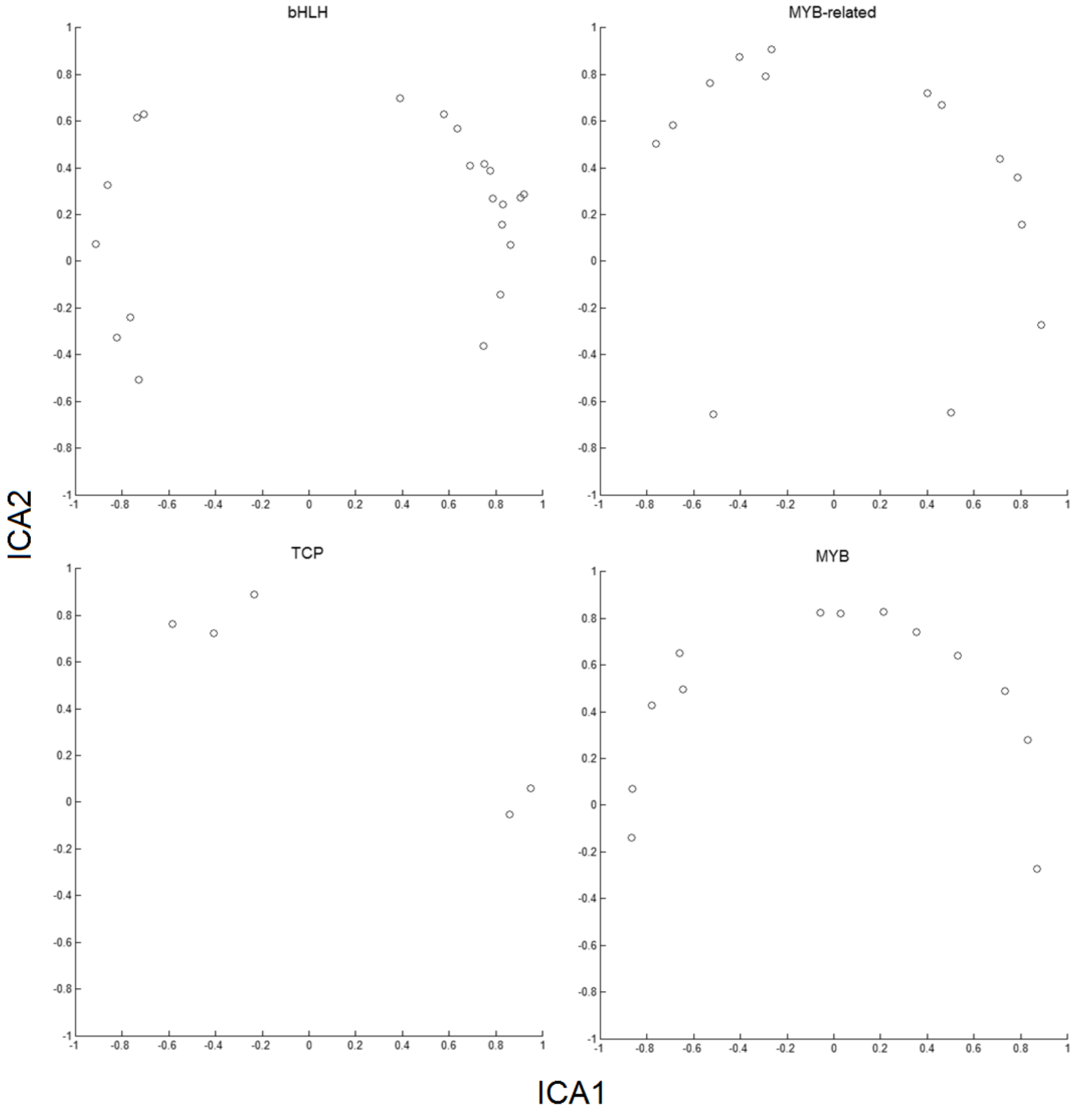


Figure 6.1: The distribution of TF with respect to the circadian time (with regards to the 2 cycling components). Predominant expression of the TF family and its over-represention during different times of day are visible.

information may be suitable). Figure 6.3 shows the varying results of correlation and mutual information on detection of relationships. Figure 6.3 shows how correlation might not pick up certain real interactions versus cases where mutual information likely will. Mutual information here detects the associations that by virtue of the design can be assumed to be causative. For a given network inference method, whether some types of edges of motifs are systematically predicted less (or more) reliably than expected, it was shown that correlation and mutual information methods are best at capturing feed forward loops among other possibilities (directionality motifs, cascade motifs, fan in, fan out motifs) (Marbach et al., 2010).

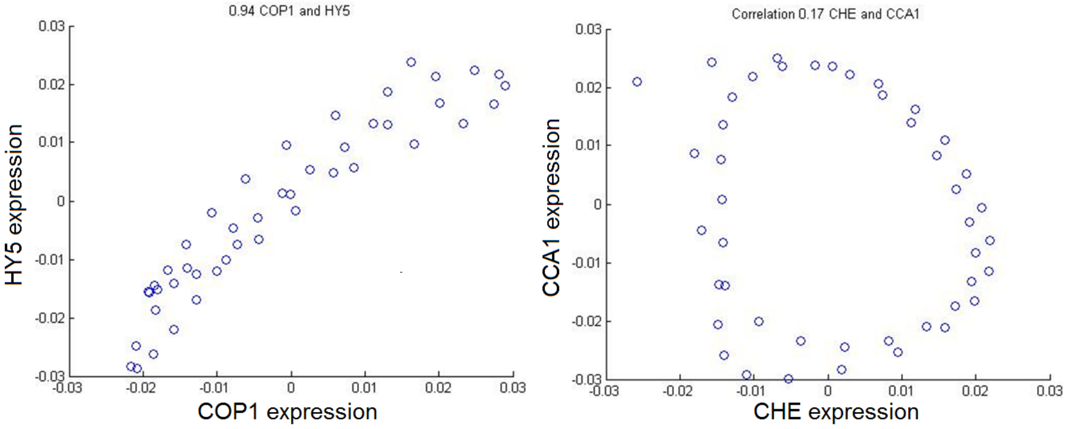


FIGURE 6.2: Scatter plot of TF:TG interaction. Two examples of diverse relationships between transcription factors and their target are demonstrated. The dependency of measure used to capture the relationship is clearly important. One example is linear and non linear. Candidate TF:TG gene relationships are assumed more likely if the expression of TF and TG are mutually dependent in at least a subset of the examined experiments. Data is derived from publicly available expression GEO At.

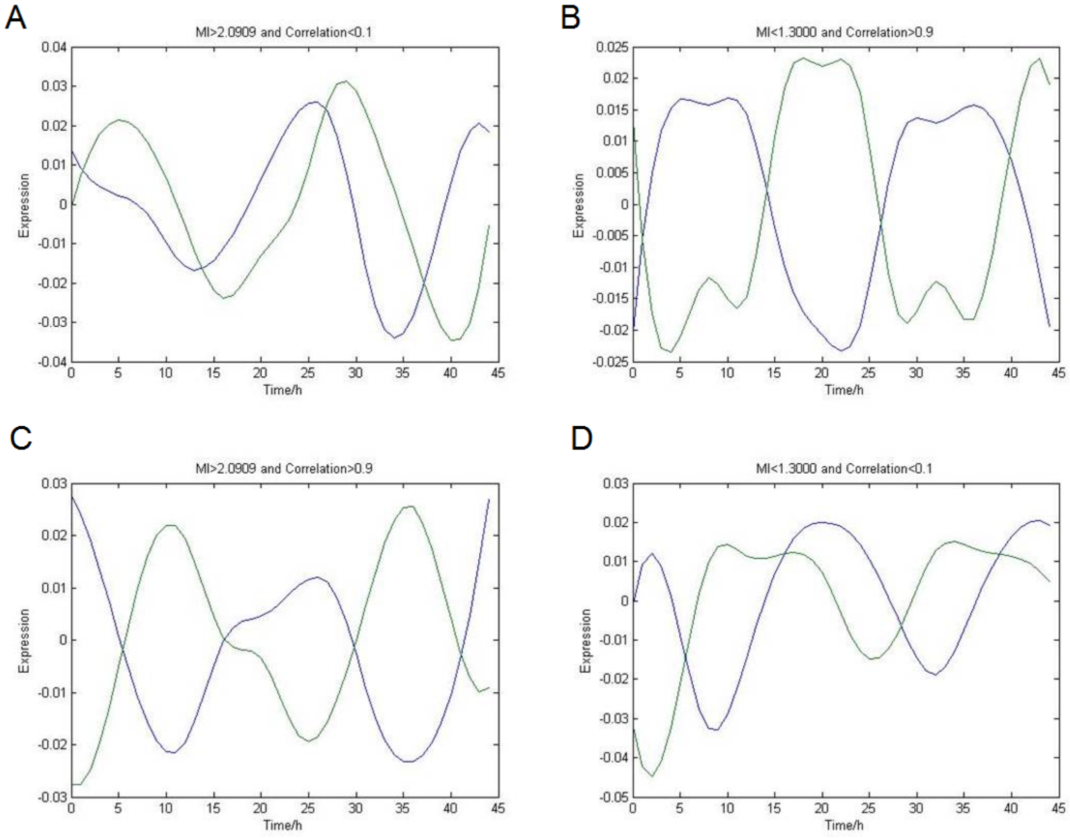


FIGURE 6.3: The potential ratios between the simple correlation and mutual information measures as depicted by synthetic plots. There are four examples of performance of MI with respect to correlation, these being the top plots, and both measures being high and both low in the bottom plots. A. represents a good MI/ bad correlation combination B. represents bad MI/ good correlation combination C. represents a case where both MI and correlation are high D. represents a case where both MI and correlation is low.

6.3.1 DETERMINATION OF THE PROBABLE TIME DELAY BETWEEN CIRCADIAN TRANSCRIPTION FACTORS AND THEIR TARGETS

Gene regulatory interactions among nodes being genes are not instantaneous, rather they are dynamic events which occur throughout a period of time. The time lag is resultant of a cascade of biochemical reactions which lead to the effect on the target by the TF. In algorithm comparison studies, best performers on all datasets are dynamic models (Lopes, 2013). Clearly, incorporating temporal information is beneficial to the inference of interactions in GRN. The performance of the adaptive lag models changes with the parameters set like the 'lmax'. The performance of fixed lag models on the other hand (lag being one time point) should be influenced by the interval length of the time series yet was shown to outperform the static no lag models and decreases the search space as opposed to scanning windows which yield high number of potential connections. The time lag between circadian TF peaks and that of their targets can range from immediate up to 12 hours and its determination is critical to the understanding of such a regulatory network. Initially a fixed time lag of 4 hours was used (the time difference between the initial change in the expression of a given regulator gene and its potential target gene), decision made upon the knowledge of circadian examples and resolution of the data. Ideally, a measure of the optimal lag period between TFs and targets calls for an independently known, unambiguous set of confirmed pairs to serve as a test set. Such a test set was obtained from confirmed interactions in literature stored in BioGRID 3.2 repository. Clearly the well studied examples of TFs and their known targets reduce the number of unknowns. Figure 6.4 portrays a histogram of time lags derived from exactly such a database focusing specifically on regulatory protein DNA interactions (the list is present in the appendix). Biogrid informs and stores confirmed interactions of different nature. The peaks for the stored input genes were derived using phaser. To a surprise the figure shows the highest frequency at no lag, with 50 percent of frequencies explained by -4 to 4 hour difference. The information within circadian biology would rather point to 1 point, 4 hour lag which is possible looking at the histogram yet not the most probable. As the histogram represents a subset of interactions further assumptions will be made based on both known circadian TF:TG pairs and the BioGrid histogram. Another alternative would be to use coexpression, combined with GO and promoter element overrepresentation, to identify tight, mutually informative clusters (with peak phase and wave form). These could then be the starting point for identification of candidate TF, on criteria of the wave form and time gap. As cis elements are not clear cut defined as chapter 4 results have shown and it is rather the objective here to define the global scale, that approach was not carried out per se yet was indeed carried in a reverse order, that is integration of GO, semnatic similarity to inspect the obtained

TF clusters and extraction of cis elements for RF predictions. Figure 6.5 portrays the resulting distribution of the number of targets in the downstream TFs groups reliant upon the selected time lag, obtained from the real data used to create the circadian GRN. The groups are the circadian transcription factors and the size is determined by the number of targets per each. That is of relevance as small groups might be interesting in terms of targets yet they cannot be used to extract *cis* elements due to group size limitations. Such a methodology likely reflects the differences between TFs with respect to time delays in their translation and the trans-activation of their target *cis*-regulatory elements. A method of time lag estimation based on confirmed data no doubt increases the accuracy of predicting gene regulatory networks and was derived from real world examples and analysis of distributions of known genes. A similar approach has been evaluated using time-series expression data measured during the yeast cell cycle. Clearly, one is more likely to observe a significant statistical correlation between the expression of a regulator and its target if biologically relevant time slices are used (Schmitt et al., 2004, Zou and Conzen, 2005).

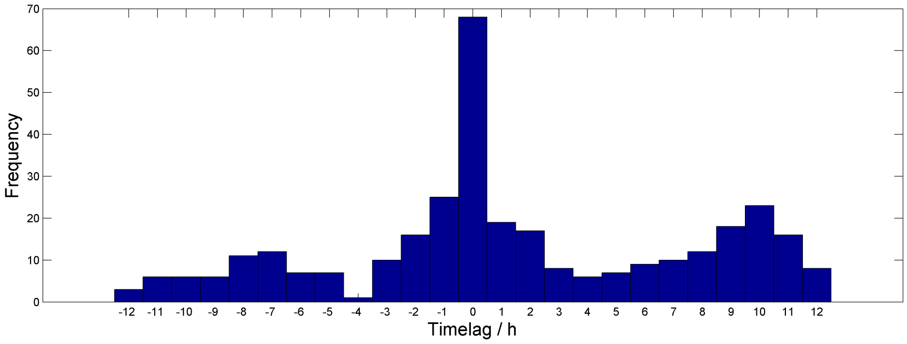


FIGURE 6.4: Frequency distribution of timelags given a set of confirmed regulatory, protein DNA interactions. These interactions were derived from BioGrid database. From 12114 interactions derived from BioGrid, 331 are portrayed on the graph. 171 interactions fall within -4 / 4h lag period constituting over 50 percent of the total set.

We found the potential targets of the circadian TFs giving rise to output groups. We confirm the consistency of outputs in the 916 project set which was a self assembled dataset encompassing all the public Arabidopsis experiments available at the time to the best of knowledge. It filtered the noise out as the objective was getting greater precision at the expense of recall (false negatives were not the danger here). This data was compared against the output from IOTA stands for the Inner Composition Alignment for inferring directed networks from short time series (Hempel et al., 2011). It is a permutation-based asymmetric association measure to detect regulatory links. It has several advantages including the ability to identify coupling and its directionality, the ability to distinguish direct from indirect couplings, ability to infer autoregulation (for example adaptive mechanism by which a subsystem regulates itself), its applicability

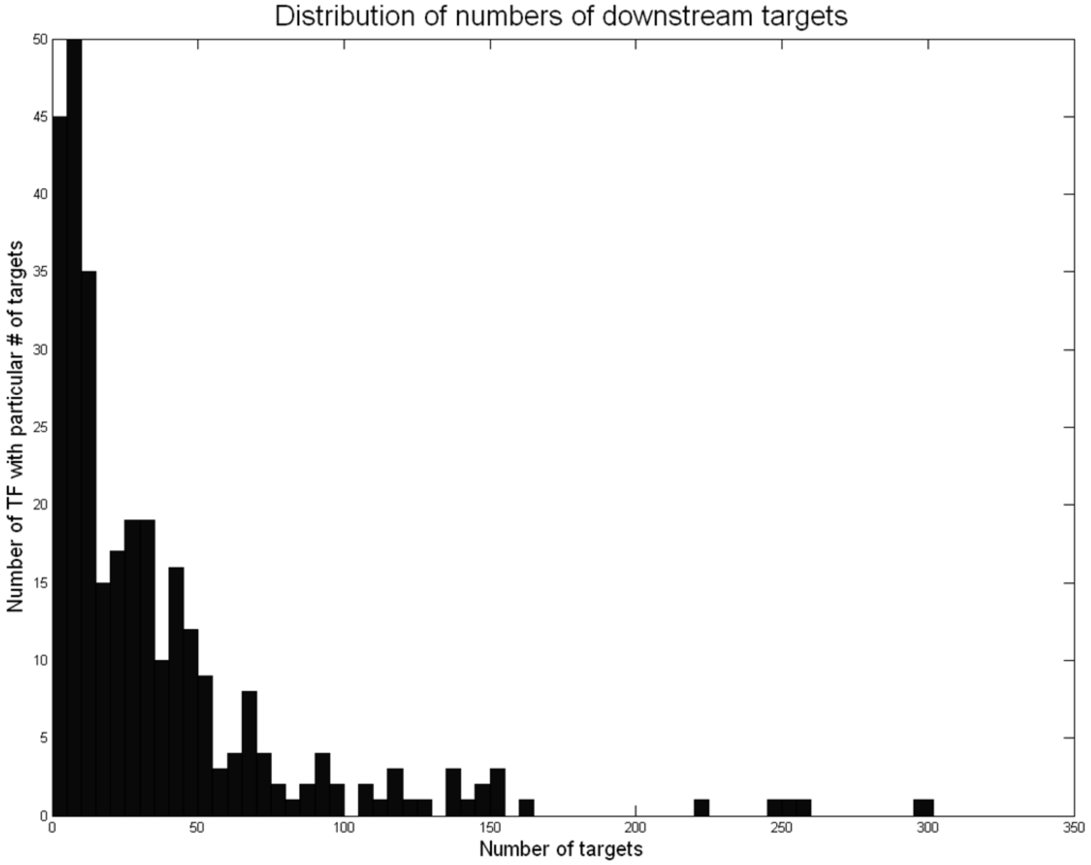


Figure 6.5: The size of the TF target groups given the time lag. There is a clear association with the majority of TF having small number of targets, and solely several outliers having large target group sizes of for example 300. This is real circadian data with 4 hour time lag.

to short time series, and no explicit dependence upon time. In essence if G is a TF and Y and Z are the outputs we used our method to find the subset Y and Z and we validated that subset against the 916 project described in the methods section (2.2.13). Then essentially we had all the data to reconstruct the regulatory network using a vector of features derived on individual TF groups that were larger than 20 genes (which is the minimum number of motifs to extract credible elements). Figure 6.6 portrays the initial searches for targets of TOC1 where the output could potentially be at the time of expression that is fixed at time 0 which does not seem biologically realistic as seen in the top A subgraph. Here TOC1 is in red and the likely output cluster is the negative correlation cluster. Subgraph B allows for delayed time lag of 8 hours. Subsequent modelling operated on 4 hour time lag and gave rise to output groups like that seen on Figure 6.7. It shows the target cluster for a given gene SI3 at one timepoint lag period.

To make sure the results are recovering real interaction one of the most promising methods GENIE3 based upon regression and RF was additionally used here to generate gene regulatory networks for TFs and their targets. It is characterized by the ability to deal

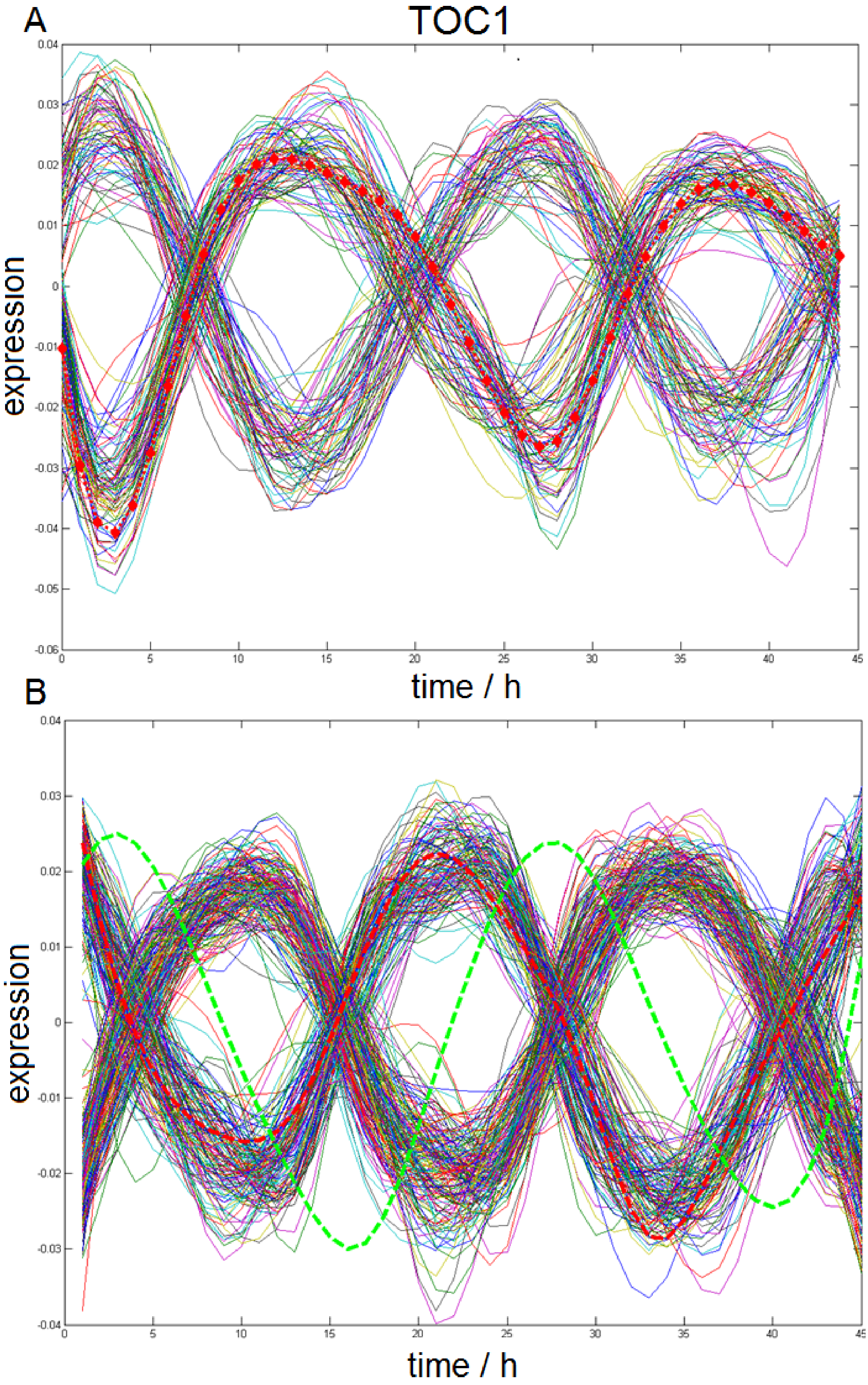


FIGURE 6.6: The example output for TOC1 occuring, at the same time in the A plot (red line) and occuring at a 8 hour time lag in plot B which is the lag with greater biological validity.

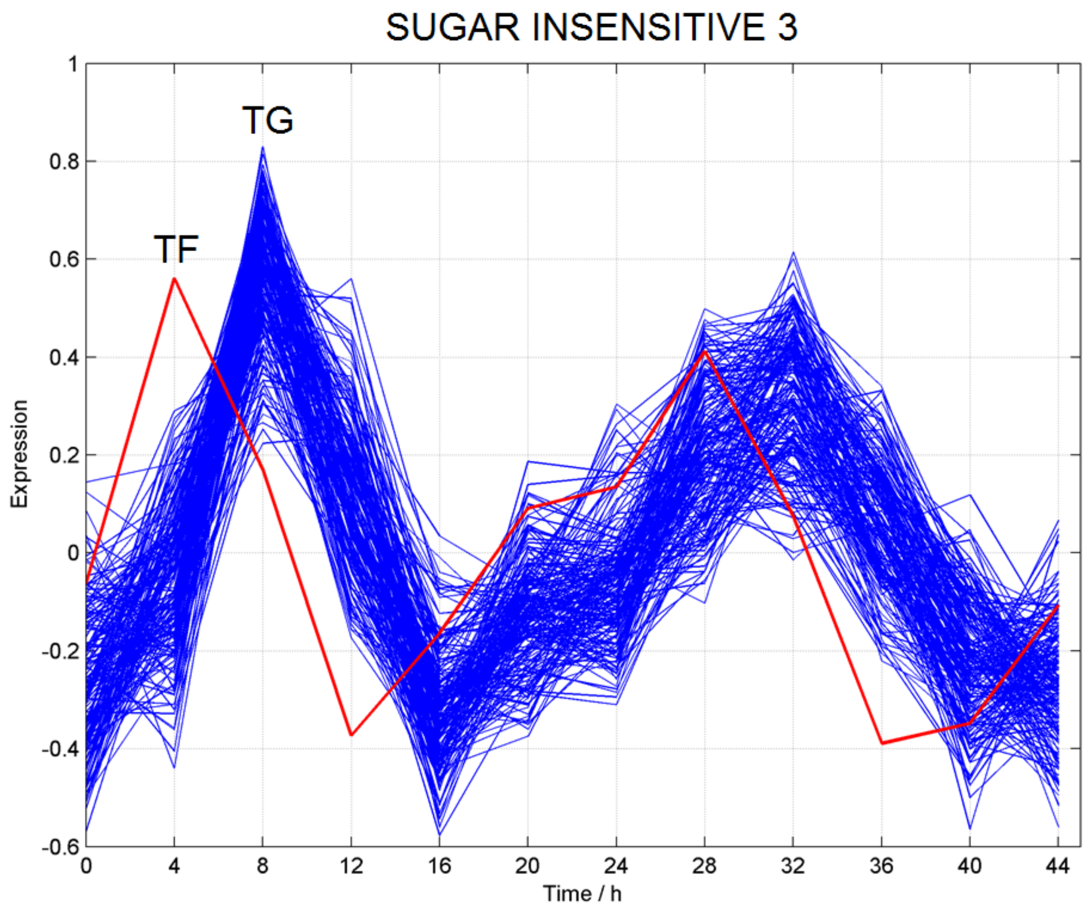


Figure 6.7: The output of TF:TG interaction occurring at 1 time point 4 hour lag period using the method of choice. The blue cluster is the TG cluster whereas the red plot is the TF itself

with combinatorial and non-linear interactions, ability to infer directed GRNs and scalability. To generate the network, driver genes were classified as TF genes and target genes. Although 1000 interactions were inferred for each of the group, an interaction score¿0.037 based on empirical distribution and constituting 99.9 percentile of all the connections was chosen as a cut-off value. Using this information the regulatory networks were independently inferred for circadian targets of each TF. The TF and target gene expression file together with output are in the appendix. All 3 methods: time lagged correlation, MI derived interactions and regression RF GENIE3 interactions were carefully examined. As integration of several methods seems superior solely the interactions present in at least 2 of the 3 methods seem the most plausible.

6.3.2 IDENTIFICATION OF POTENTIAL TARGETS, THEIR *CIS* ELEMENTS AND THEIR EFFECTS UPON THE PHIs AND THE PSIs

The regulatory TF:TG network have three types of output: targets per TF, cis motifs per TF (which allow for classification of the TFs) and the regulatory network itself. Whereas all three components are of relevance each drives hypotheses of different type like the specific connections for a TF, the cis motifs worth looking at and the topology and hubs within a network for example. The entire set of regulators is present in the appendix together yet two interesting instances one new in circadian biology and one established are presented in the text. One such instance is of TF RVE7. Its thirteen candidate targets include: HFR1, P1R3, HYR1, 2 zinc finger family proteins and Oxygen-evolving enhancer protein 1-1. The overreprsented GOTERM is reponse to light intesity. Among the overrepresented elements are known motifs like the WBOX AGTGACTA, GTAC element AGTGTACG and a unknown element ATCCCG. RVE7 is itself at target of PIF4. Such a informative set of information is useful for of course novel genes in the circadian field yet double checking with old cases like TOC1 is most certainly worthwhile. The group of targets includes CCA1,ABA1,CDF1 and CDF2, SNARElike portein and several others. Interestingly the overrepresented GO terms include: response to radiation, response to light, response to abiotic stimulus, regulation of timing of transition from vegetative to reproductive phase (including Dof proteins) and regulation of meristem development. The overrepresented elements include: GTA1 motif from the known list and unknown group of elements. Itself it is a potential target of LHY and CDF1. Both CCA1 and LHY interactions with TOC1 are annotated as genetic interactions in BioGrid. Such information is available for all the circadian TFs in the appendix. The next step is GRN reconstruction was the focus on cis motifs. These were extracted using ELEMENT tool described in Chapter 2, section 2.2.10. The classification was RF using leave-one-out cross-validation procedure. Figure 6.8 shows the *cis* grammar per target group of LHY gene. Specifically, it shows the changing contribution of motifs given the target genes. Such output was then repeated adding one more feature; that is, frequency for example TF in Figure 6.9. In Figure 6.9 the combination of motifs like ACATAC, the EE and the AACGTG matters. This particular figure focuses on the targets of MYB20, UNE10, WR14.

For each TF that is called circadian and adhered to a group big enough, a set of information as shown in Figure 6.9 was presented (the entire set is in appendix). Here we can see the differences in target group sizes, number of important motifs and their frequency where frequency seems to be more relevant factor than position. Figure 6.10 shows the ordering of all the circadian TF through their angular position. Such ordering reflects

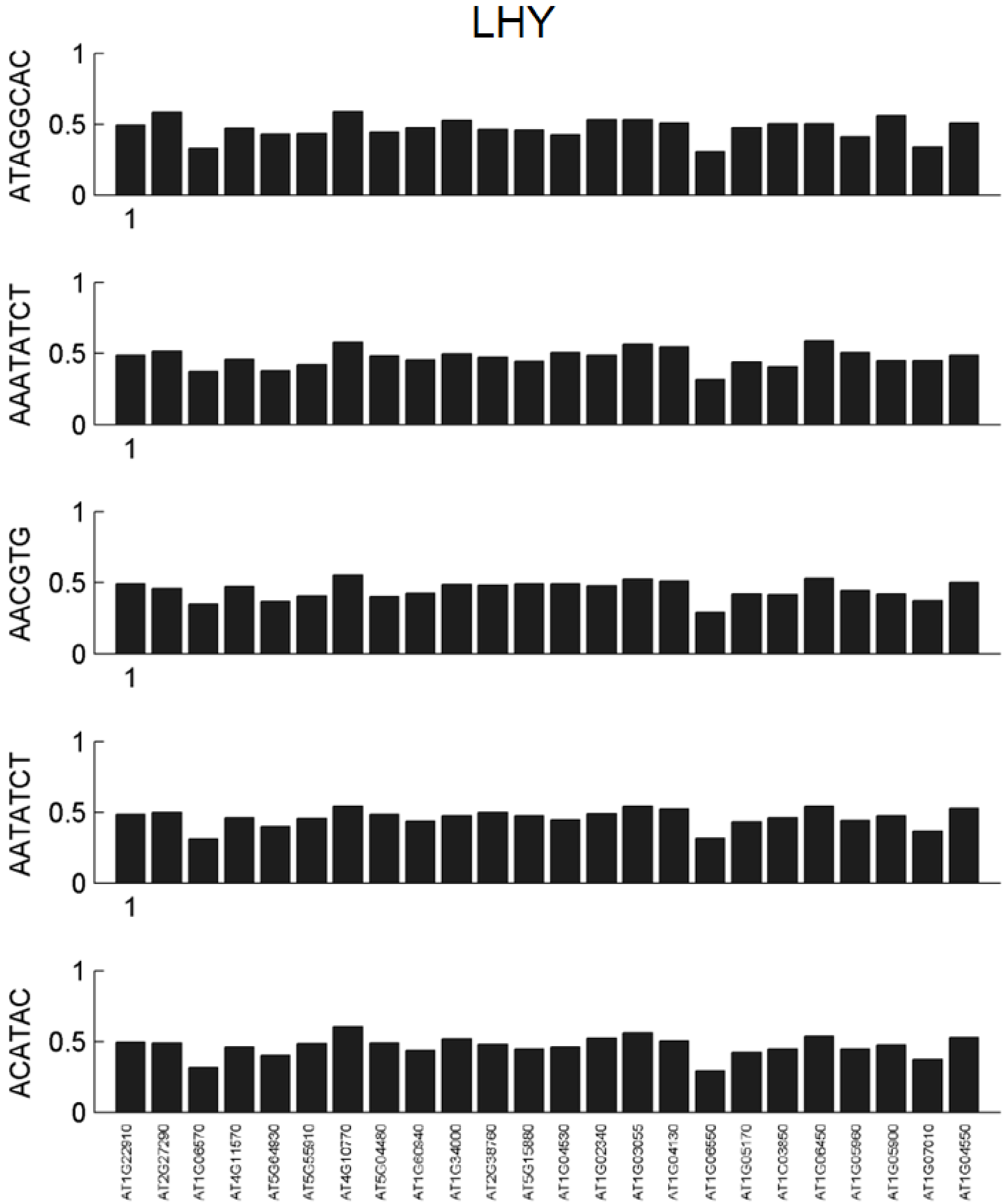


FIGURE 6.8: The output *cis* elements as derived from the TF target group.Each bar represents a target of LHY with the change in contribution per gene given the motif. It is a striking combination of elements with 2 EEs being present. Such information is showing the dominant regulatory locations, the hallmark of a specific target group.

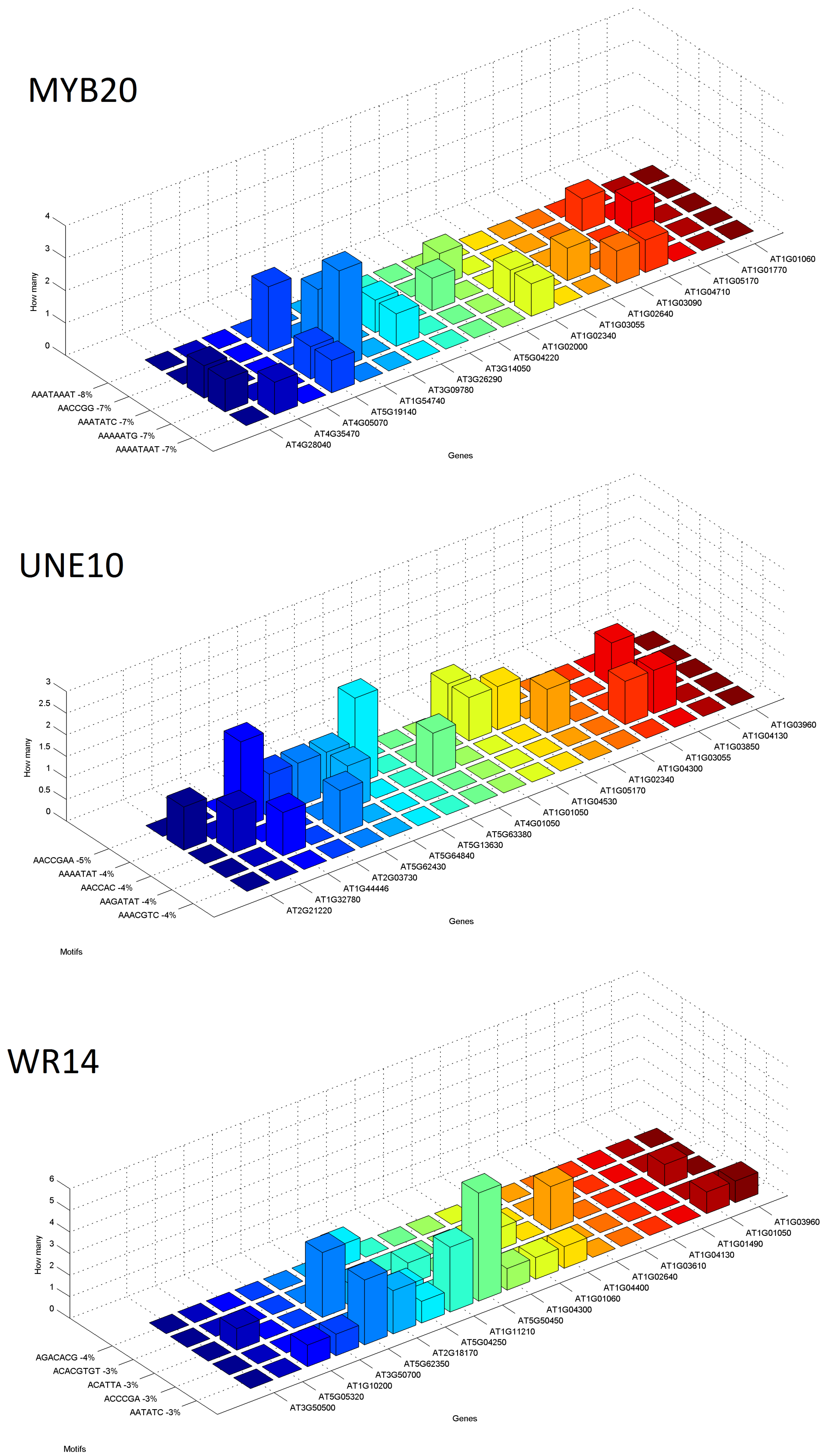


FIGURE 6.9: The concept of motif grammar for multiple circadian TFs. The figure portrays example TFs with their motif grammer, that is promoter design.

the distribution of TF in circadian time. It is interesting to see how this ordering changes in terms of phases when the condition/mutant would be in the altered state given such data will be available. The *cis* grammar, together with TF phases state became the building blocks for the regulatory networks seen in Figure 6.11. Now given the ordering and knowing the grammar, the similarities between TF were extracted. Doing the same with regards to the *fhy3* could show reordering of TF and targets in temporal terms. Figure 6.11 shows the links between top TF and cis motifs. Among the hubs motifs are TGAAAA, AATATCT, AAAAATC, AAAAATG and ATATTAA.

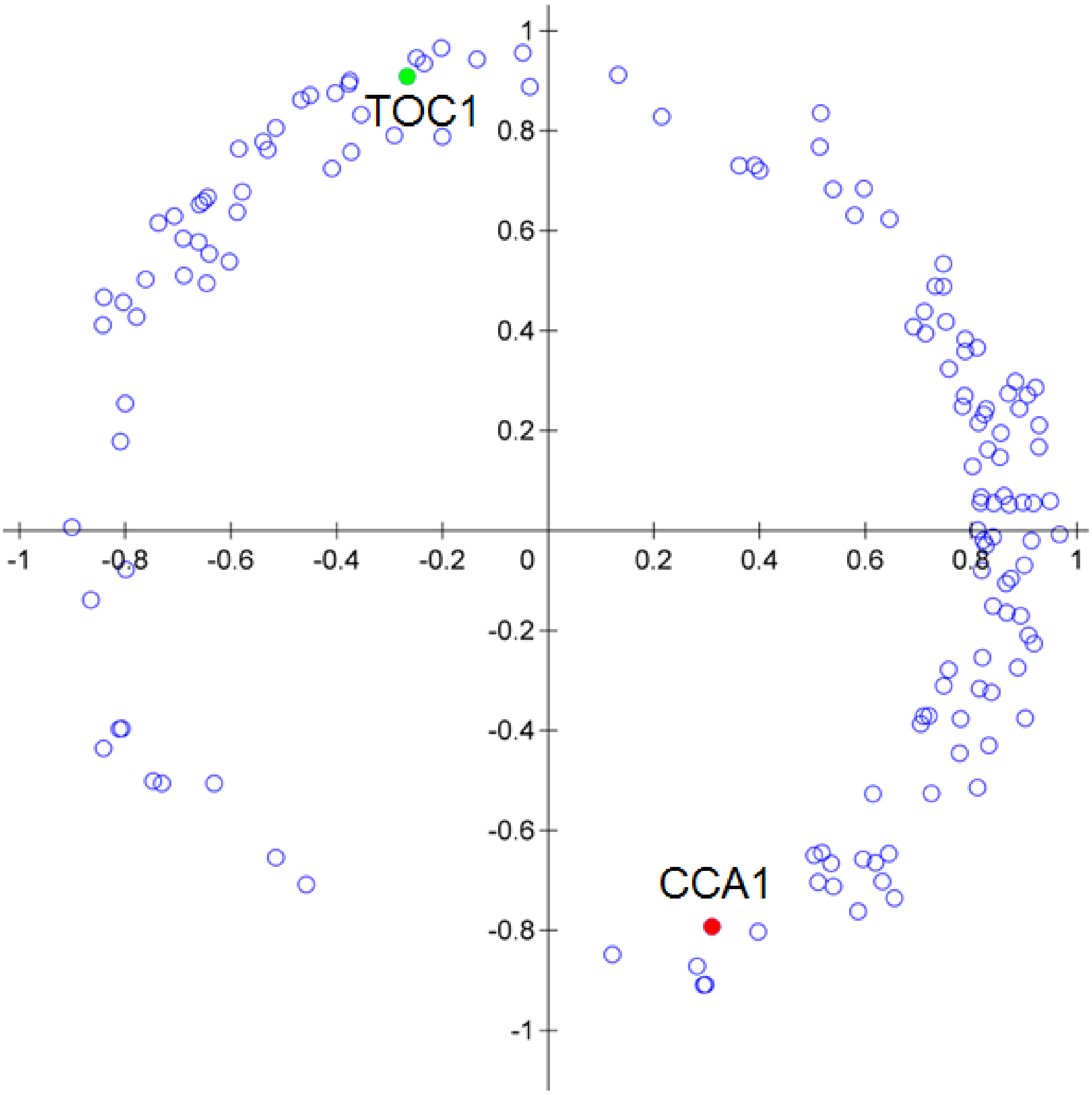


FIGURE 6.10: The global ordering of the circadian TFs looking at their angular positions. The red and green colours represent *CCA1* and *TOC1* presented here as a reference guide. We conducted a comprehensive screen and statistical analysis of gene expression to work out a gene network comprising at least 5 core elements.

In the present analysis we identified a set of TFs and their target groups of genes including core clock and output genes like *ELF4* part of the evening complex and *PIF4* implicated in the photoperiodic control of plant growth. Among the targets of the *PIF4* are the auxin associated genes, giberellic acids, ethylene and cytokinin genes. These 2 genes will serve as examples as after all this experimentation we can see all their networks at once as seen in Figures 6.11, 6.12 and 6.13. The knowledge of angular position is not important in itself but it is indeed important with reference to changing states which could be for example *fhy3far1* double mutant. Nevertheless, the TF circadian reconstruction led to many potential findings that were validated computationally and call for experimental validation. Such a comprehensive model has TF acting at different phases with their regulators reveals the *cis* regulatory code responsible for particular phases. That leads to construction of phase vectors as done in mammalian clock studies, that should allow for exciting modifications (direction and length of a vector represent the phase and the amplitude of the wave). Phases describe the distinct stages, building blocks of a process. They are characterized by the angular displacement between a periodic quantity and a reference angle. That phase angle displacement is the key to deciphering effects of particular mutants, stimuli which can for example induce phase shifts (displacements of oscillation along the time axis). It is not solely the inducing component yet additionally the phase at which the stimuli for example is presented that affects the phase response curve. Having the ordering of when does a TF peak on a plot we see how certain TF act in concert (we can hypothesize at least that they could theoretically), how certain ones are similar to each other like *CCA1* and *LHY* and how knowing the phases and the *cis* grammar, one can reconstruct those TF to achieve a desired phase. That is particularly nicely portrayed on a Cartesian plot with circadian hours. One can foresee that a quantitative understanding of gene expression output will call for integration of many inputs: transcription factor levels and binding constants, competition and cooperation between transcription factors, transcription factor dynamics, network structure, and nucleosome occupancy. That would likely eliminate many false positives. This study describes a systematic method to discover previously unidentified roles for putative TFs involved in circadian physiology in plants. Figure 6.11 portrays the gene regulatory network of circadian oscillating TFs and their motifs. Our computational analysis of gene regulatory networks revealed a number of transcription factors (TFs) that mediate novel circadian functions (are interwined in not yet described pathways) at the same time as the less described one, for example *RVE7* and *RVE8*. There are solely a few transcription factors (TFs) within the core circadian circuit (present loop model genes), from this we can conclude that, circadian oscillations in gene expression in different tissues are unlikely under the direct regulation of known core circadian TFs, but rather they are likely to be regulated by other TFs that are themselves regulated by core circadian genes and this data is serving to answer these questions. Previously major breakthroughs in the mammalian clock were made by small networks provided

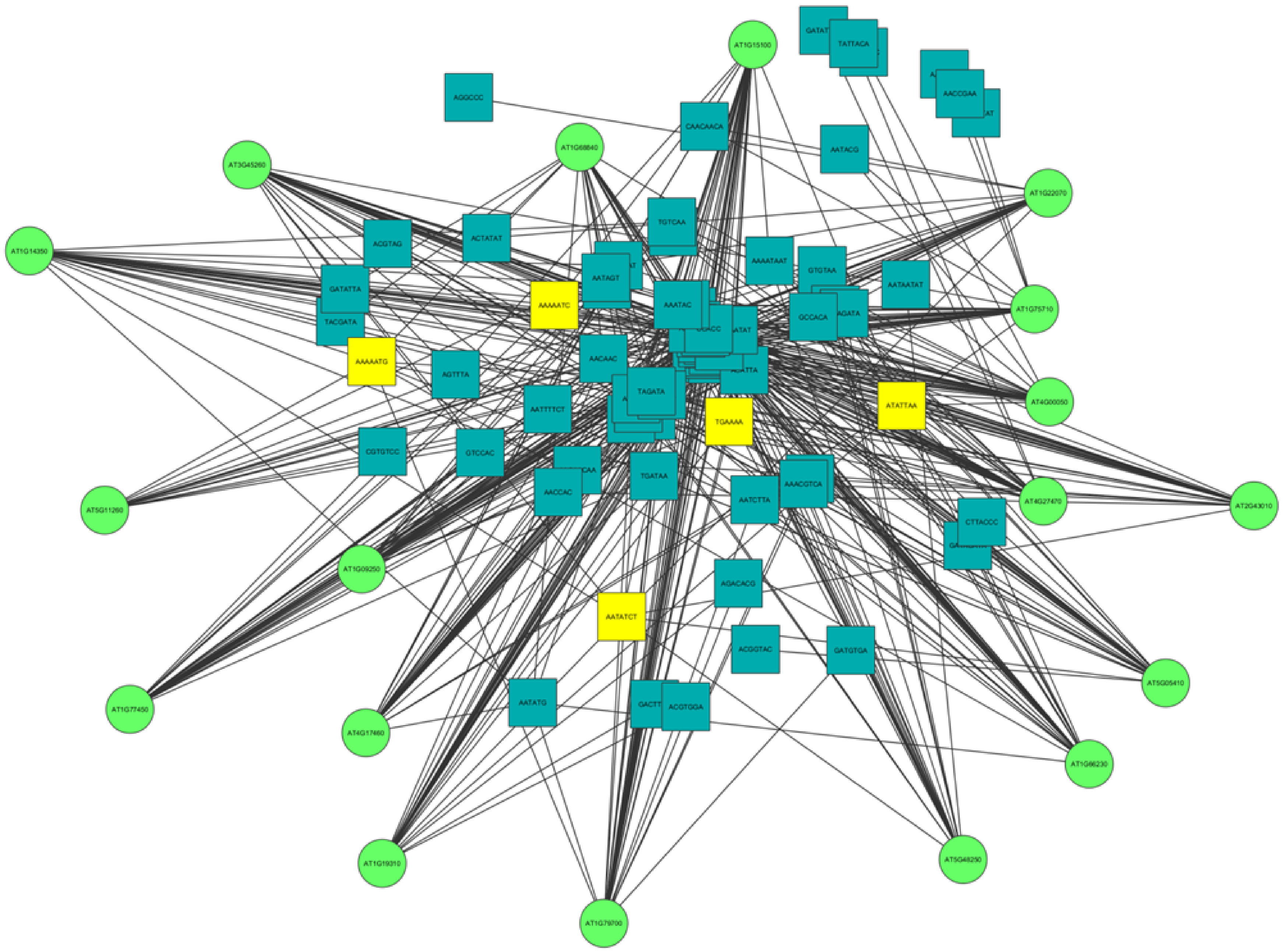


FIGURE 6.11: Predicted regulatory events between all circadian oscillating TFs (given the size of the group) are connected by TF motif relationships. Circles represent circadian oscillating TFs and boxes represent *cis*-regulatory elements characterized by TF DNA-binding motifs. The green nodes are the TF, the blue nodes are the common *cis* elements whereas the yellow nodes are the top hub motifs.

by (Ukai-Tadenuma et al., 2008). These authors constructed a small-scale gene regulatory network consisting of 16 genes and 3 *cis* regulatory elements based on in vitro luciferase reporter assays(Ukai-Tadenuma et al., 2008). This study presents the first circadian plant gene regulatory network on the system level. This is the grand scale in plants. The future incoporation of knockout and mutant microarray experiment results with promoter sequence analysis can greatly facilitate the identification of functional transcription factor binding sites. That has been seen on a small scale portraying the methodology. This explorative unsupervised methodology will be further discussed in the next section. Specifically in terms of the network progression like shown in Figure 6.12 and 6.13 where the rewiring of core clock genes in terms of circadian time is seen. These figures portray the static interactions whereas the dynamic interactions are present in Figure 6.14. The central hubs are in yellow whereas the connections between them are day specific annotated as blue edges and night specific annotated as red edges. These key genes include PIF4, RVE7, RVE8, CCA1, TOC1, GI and PRR7, ELF4 and LHY. RVE7 is capable of its own suppression. Here the specificty of targets demarcates its involvement. 6.14 The differential TF modules in Figure 6.14 contain some interesting characteristics per group. For example the CCA1 day specific regulators are overrepresented in GO responses to osmotic stress, salt stress and stimuli, and to name a few genes include LHCB4.3, CYP98A3, SEC61 among many more for which detailed list is provided. The night list is 7 times small and contains mainly yet undescribed targets. For TOC1 night connections among several genes include a interesting link with ELF9 involved in negative regulation of flowering time. PIF4 has a night specific TOC1 connection (confirmed via BioGRID) and a CRY1 day and zinc finger JACKDAW. Among the overrepresented GO day processes is regulation of growth, development and meristem growth. The example waveforms for differential regulation is present in appendix.

We measured prediction accuracy by the Area Uunder Curve (AUC) measure for the TFs in the regulatory network 0.79 (shown in Figure 6.15). Other measures, including sensitivity, specificity, precision, Matthews correlation coefficient and F1 score were applicable yet, in contrast to AUC, these measures called for the selection of a threshold that transforms edge weights into interactions and non-interactions, essentially drawing a point on the ROC curve. That usually raises further questions of how (at what point on the ROC curve) to define the threshold. The AUC allows for the unbiased comparison without the urge to optimise a threshold. The obtained AUC of 0.79 seen in Figure 6.15 shows the average ROC curve for the classification of TF:TG network, this being, suggesting high signal in the choice of targets reliant upon the selected cis elements. This is not to say that we dont have false positives and false negative yet is reassuring to see some common cis elements, taegets of known TF, meaning that one draw conclusions from the previously unknown proteins.

6.4 DISCUSSION

Given a network what comes next? These lead to causal maps, perturbation experiments, networks as biomarkers and networks for comparative analysis of evolution of a system (Emmert-Streib et al., 2014). Information about temporal changes in the network structure are important to understand as shown for example by (Yosef, 2013). The validation is the critical step if one is to rely on the complexity of the derived model. It calls for the use of distance functions to measure the closeness of two networks that has been done in earlier chapters. On the other hand scientific validation involves evaluating the concordance of predictions made from the model and the corresponding experimental outcomes. The former involves justification using performance of the algorithm on a shuffled dataset, whereas the latter calls for reductionist confirmations. Network models provide quantitative knowledge concerning gene regulation and, from a translational perspective, they give a foundation for the mathematical analyses leading to system-based optimal modification strategies. Gene coexpression networks can contribute to the discovery of novel functional relationships, leading to key biological hypotheses. The available experimental and theoretical research has suggested that gene regulatory networks are often times characterized with scale-free properties (Albert, 2005, 2007). Nevertheless, a mathematical model alone does not constitute a scientific theory. Here the circadian model is strictly predictive. It has a formal structure that should lead to experimental predictions in the sense that there are relations between the variables and the already observed phenomena such that the experimental observations are in accord with the predicted values of the variables. In the example of full circadian network inference, the inference procedure is a mapping from a space of samples to a space of networks and it must be evaluated as such. It is worth mentioning that the data covered up two days within one stage in the development of Arabidopsis thaliana at constant light conditions, so the results could vary at a different stages and tissues. As has been shown recently by (Endo, 2014), the tissue-specific clocks show asymmetric coupling. That is the clock is not uncoupled, rather there is a hierarchical organization with vasculature cells. In example it was shown that the vasculature and mesophyll clocks asymmetrically affect one another and that disruption of solely the vasculature cells disruption of the circadian clock in the vasculature cells affects the timing of flower production (Endo, 2014).

At present there is a recession in terms of investigations into validation, the net result being that biologists are uncertain of the status of networks proposed in the literature. That in the light of even BioGRID associated confirmations is rapidly changing. Here the specific network can be approached from 2 angles being the validation of interaction and the validation of particular *cis* elements. Hence clearly it is the validation with a variety of in silico and in vivo datasets that yields the best outcomes. Certainly regarding the presented network there is noise resulting from the additional influences which may have arisen in the protein and/or gene space including post-translational and epigenetic effects. The inference of Gene Regulatory Networks from such a time course data has proven useful for disentangling relationships between genes to the extent that one can now focus on the most important TF per phase, causing for example transcriptional delays and the rate limiting effects used to mimic the chemical interactions (Kholodenko et al., 2012, 2002). In independent previous studies all algorithms identified *ELF4* as a target of *LHY/CCA1*, consistent with previous modeling using *ELF4* with the clock genes. A few inferred networks had *ELF4* feeding back into the clock, several methods, including nonstationary suggested it was a terminal node in the network (Penfold et al., 2012). Here it was shown with its target group not in terms of its links in the entire network yet rather in a manner focused on the design of the potential phase vectors. The overlay of multiple data types with multiple tissue types will certainly lead to the understanding of the plant chronome, far beyond the present understanding of individual components of it. Figure 6.12 portrays the core clock genes around the circadian clock. One example (derived from the present analysis) of a TF detected to be a night hub is *RVE8*. Genes regulated by *RVE8* are enriched for two complementary phases, with the *RVE8* induced genes (*TOC1*, *ELF4*,) enriched for the evening phase and the RVE8-repressed genes enriched for a morning phase (*CCA1, LHY, RVE8*) as described by Harmer. Such dynamic network progression across six time phases was recapitulated using the data used with additional links added. This night hub TF that has both activator and repressor capability was central to the static analysis. The differential night network hub is *RVE7*. That is additionally a driver node directly linked to another driver node *PIL6*. Now knowing the central *cis* motifs in a TF network like those presented in Figure 6.11 together with central drivers nodes is the crux point for understanding yet the starting point for alteration. Testing the methodology involves counting the number of links correctly predicted by the algorithm (true positives, TP), the number of incorrectly predicted links (false positives, FP), the number of true links missed in the inferred network (false negatives, FN) and the number of correctly identified non to be links (true negatives, TN). Here this was carried out from the perspective of motifs found and predicting classes of TF. Apart from applying RF promoters screening as presented in Chapter 4, incorporating CHIPseq and one-hybrid system will further put hard constraints showing which transcription factor can bind target genes allowing to verify key hubs. CHIPseq allows to elucidate fine binding sequence of a TF and at the same time describes exact position on a genome that is being bound. Figure 6.13 is another potent example of the route for alteration of *RVE7* and its collaborating genes to control pathway. *RVE7* is interwined in a control path including *SEX4*, *CCA1*, *RVE8*, *RVE7*. It is part of module 2. Both the differential pathways yield over-representation terms suggesting its core circadian function.

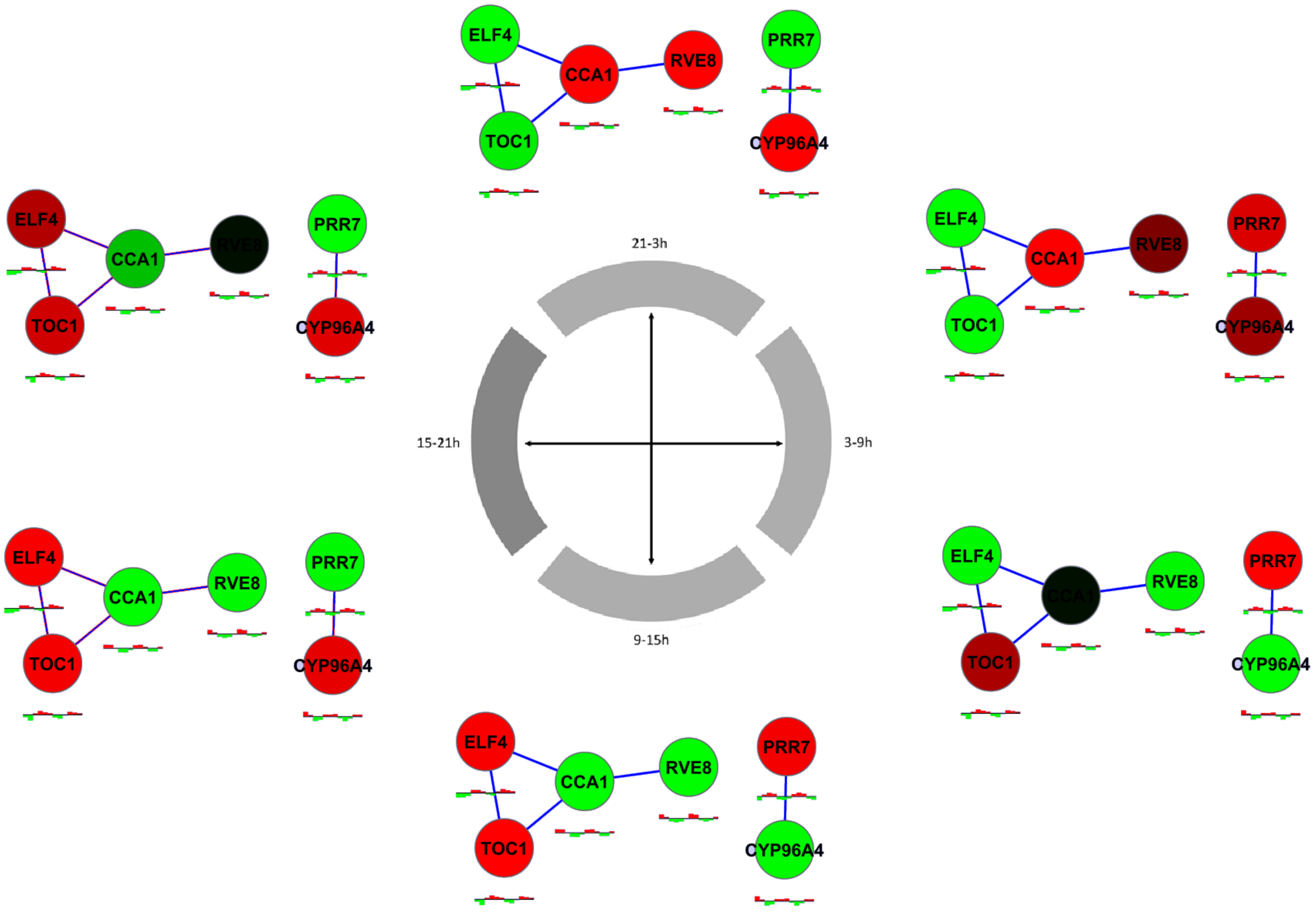


Figure 6.12: TF network progression across the six time phases. The colour range represents the level of expression on a gradient with green-low, red-high for the nodes. The phases for clock regulated transcription factors span the entire 24h range.

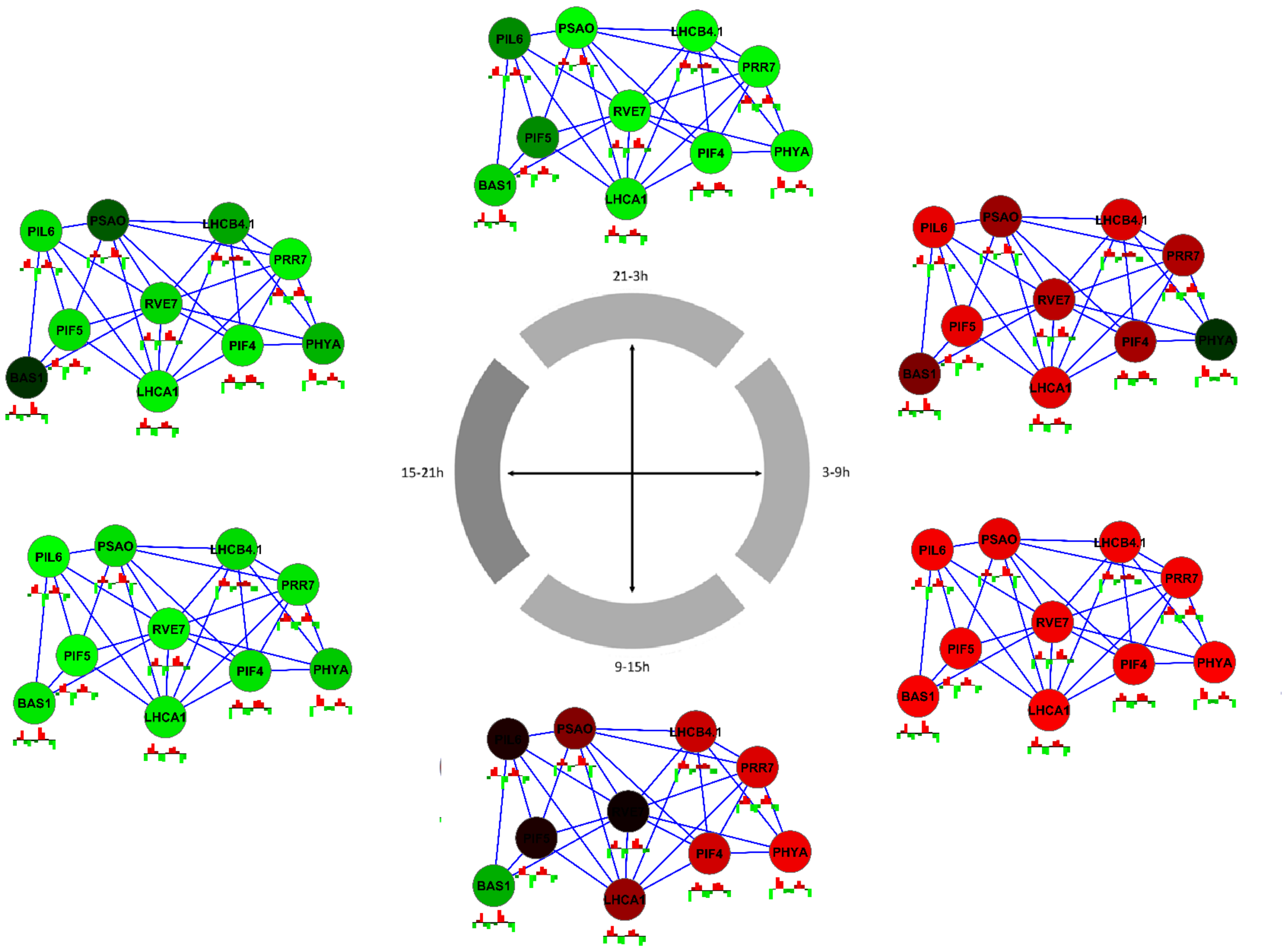


FIGURE 6.13: *RVE7* interactions as they occur in circadian time. The links constitute thresholded interaction. The colour range represents the level of expression on a gradient with green-low, red-high for the nodes. The phases for clock regulated transcription factors span the entire 24h range. The pattern of colour changes across time is relevant rather than the particular gene names.

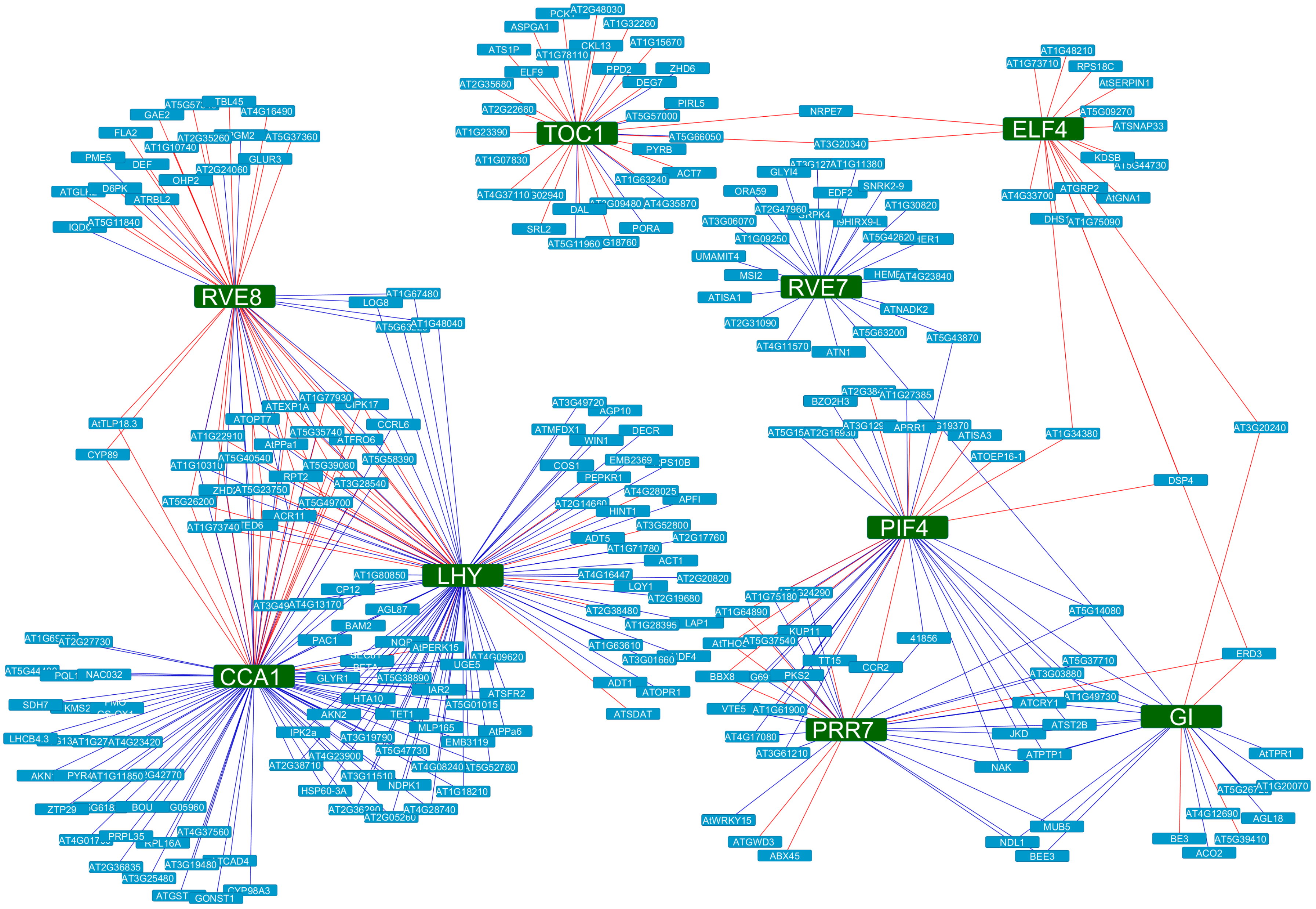


Figure 6.14: Differential day and night interactions with day links annotated in blue and night links annotated in red. The interactions are among 9 TF annotated in green. These 9 TF form modules with their targets. The specific details are provided in the text, hence the here the pattern and the main day and night connection are interesting

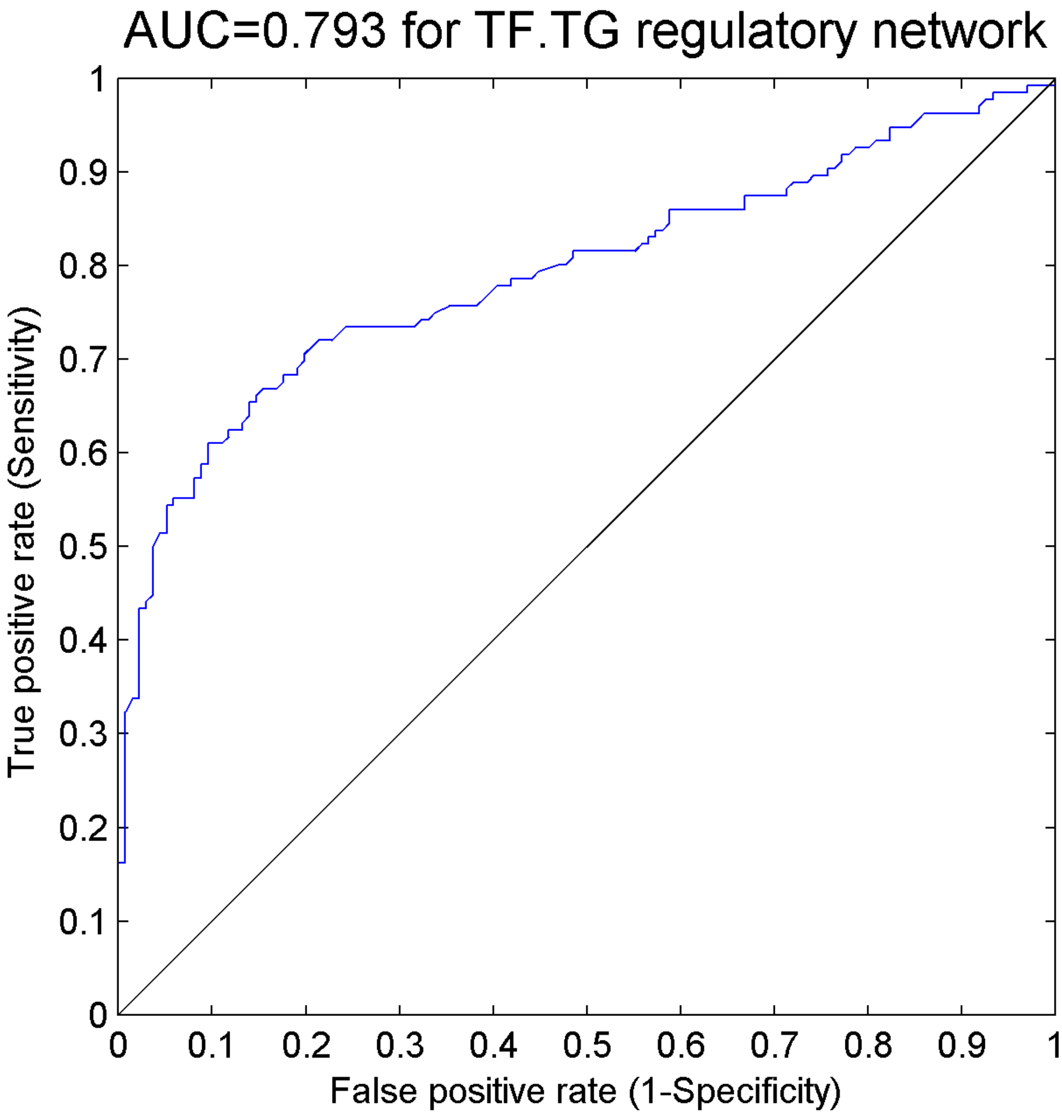


Figure 6.15: AUC for TF:TG regulatory network. The AUC is 0.793.

Chapter 7

THE MANIFESTATION OF LIFE RULED BY THE CIRCADIAN CLOCK

7.1 DISCUSSION

7.1.1 FINDINGS

The research work presented in this thesis occurred at the edge of scientific revolution, where the paradigm shift in Kuhnian terms occurred. It is the increasing popularity of complex network research and systems biology that allows for a holistic, deeper understanding of circadian clocks. Specifically, this work presented a comprehensive analysis of the circadian clock per se, that is, looking at wild type and mutant datasets. The main actors, as in hubs, and the main communities, here modules, were delineated, showing the four circadian modules preserved across data replicates and species. The analysis of differential interactions together with groups of genes allowed for higher order inferences regarding the logic of operations giving rise to *cis* elements and allowing the construction of synthetic promoters. That was further complemented by a regulatory network based on transcription factors, highlighting the potential candidates for thorough confirmation. The main findings of this work can be discussed separately, corresponding to chapters. First, the identity and nature of the circadian system components that play a significant role in generating the oscillations under study were determined through presentation of a novel method of detection of circadian genes. The "elements list" of the system was determined and validated against existed elements. Secondly, the network of interactions between these components was mapped out, and their natures determined;

and, lastly, it was necessary to use this information to derive the understanding of how the dynamics of the system emerge from the underlying interactions network.

In conclusion, we recall the main contributions of this dissertation. Essentially the work involved module discovery leading to identification and analysis of different module sets including the active circadian modules in a given condition, the conserved circadian modules across species, the differential circadian modules and composite circadian modules. That was achieved through the use of several algorithms and novel methods operating on several core concepts; these being, biclustering approaches, network propagation methods, exact methods, simulated annealing, genetic algorithms and greedy algorithms (Ideker and Krogan, 2012). The findings are *(*sensu strictu) based on generated data reflecting computation rather than presumed models. Feeding further *(*a priori) information like positional and evolutionary conservation, did improve the machine learning components of the work. The modules in operation and their preservation were inspected. Interactions were extracted on a totally novel level, which is here of particular importance; that is, with regard to time and with regard to specific mutant conditions. For the extraction of significant synthetic promoters a *cis* element prediction tool was devised all leading to the creation of a TF regulatory network. The most recent studies of transcriptional regulation can be allocated to one of the three potential categories: statistical approaches, the probabilistic approaches, which try to discover structure in the dataset as formalized by probabilistic models, and the linear network models, which aim to learn explicit parameterized models for pieces of the regulatory network by fitting to data. Here we infer the circadian regulatory network naming the core elements and the TFs from the perspective of the clock using a time lag methodology and a regression random forest algorithm GENIE3. The learned function of core *cis* elements, if verified experimentally, should allow for the first, so to name it, redesigned circadian studies. The progressive promoter element combinations allow one to classify conserved orthogonal plant circadian gene expression modules. These motif combinations change in a consistent, progressive manner from one phase module group to the next, providing for the first time a potential global description of the topology of the plant circadian system. We believe that this work will represent a crucial advancement in a promising direction. It offers potential targets for verification and provides several *cis* prediction methodologies at the same time. Furthermore, the potential exists for creation of a *cis* prediction tool based upon this work, designed for multiple conditions that will be made available through iPLANT and that will allow predictions to be made using combinations of motifs. The creation of a regulatory network and its potential points of control should be the hallmark of circadian plant network alteration. It was shown that a circadian period resonant with the period of the environment is particularly crucial for anticipation of dawn and the timing of nocturnal events. Furthermore there is short-term and transient plasticity of the period of the *Arabidopsis* circadian network. In other words certain time phases are more important than others, more permitting and allowing for uncorrected delays. The symbiosis of the circadian molecular studies with network analysis will show which phase windows are of greater relevance for example for increased immunity versus better harvest (phytochemicals). Considering the recent breakthroughs like engineering novel multigene pathways to increase photosynthesis in leaves and to recapture carbon dioxide from photorespiration, and the direct linking of photosynthetis to the clock (Dodd et al., 2005), such research becomes even more promising. That becomes particularly agronomically important in terms of understanding the vegetative yields at different circadian phases as it seems there are better and worse time for harvest collection.

7.1.2 THE SYSTEM IS > THAN THE SUM OF ITS PARTS

Time is in essence of cyclical in nature. Empirical data suggests that many scaling relationships take the form of power laws with exponents that are multiples of one quarter. Günther and Morgado proposed the 1/4 allometric exponent for biological rhythms, considering the fractal nature of biological time (Günther and Morgado, 2003). Timing is the instrumental aspect of living systems. There are several examples of chronobiological variables. The circadian rhythms are a prime example of generating such periodicity. An oscillatory network showing that timing features can be designed synthetically was successfully devised using a series of transcriptional repressors that controlled expression of GFP by Elowitz and Leibler in 2000 (Elowitz and Leibler, 2000, Sprinzak and Elowitz, 2005). To achieve the designed periodicity, Elowitz and Leibler developed a mathematical model for transcriptional/translational rates and decay rates of both mRNA and repressor proteins, and the GFP reporter. Thoroughly studied, synthetic gene circuits that control the temporal profile of gene expression can elucidate the contributions of expression dynamics to natural time-dependent processes, such as cell signaling, cell-fate determination, and development (Esvelt and Wang, 2013, Gardner et al., 2000). These various schemes have demonstrated the classic NOT, OR, NOR, and AND gates that are used to build larger logic evaluators and computations (Albert, 2007, Slusarczyk et al., 2012). Subsequent generations of improved oscillators have interlinked positive and negative transcriptional feedback loops of different strengths to drive more robust oscillatory dynamics with tunable periods and amplitudes in bacterial and mammalian cells. In the post-genomics era, the importance of precise timing and coordination of gene expression is even more apparent. The complexity can be shown in terms of a simple example of *PRR9* in *Arabidopsis* which is regulated by light, *CCA1/LHY* and the Evening Complex. Experiments like these have shown that it will be indeed possible to reprogram the timing of particular events as long as those regulatory interactions encompassing a subsystem are carefully delineated. Understanding the mechanisms behind this artificial oscillatory network could reveal important features of the mechanisms of natural circadian clocks and the development of artificial clocks in living organisms. In plants, such an inducible timing mechanism could be made to coordinate core processes like enhanced immunity and adjusted flowering time in plants. Yet the design principles have to go hand in hand with the understanding of those processes *(*in vivo), hence the understanding of these is the inevitable bottleneck.

In the most recent work in the field, the in-depth analysis of altered expression, synthetic oscillators, control systems in terms of their network properties together with tensor microarray analysis (Shaechtle, Stathis, Bromuri, Royal Holloway, work in progress) has laid the foundation for such higher order operations that may one day lead to fundamental breakthroughs in systems and computational biology. They have proven that we can determine simple causal relations independently of how complex the dimensionality of the data is relying on a statistical decomposition that flattens higher-dimensional data tensors into matrices. It should allow one to detect anomalies on a greater scale and, even better, to learn based on these anomalies. In this work, tensors, network PCA and strongly connected component analysis were the means of careful network exploration and that was reflected and demonstrated using nodetrix and portrayed in netLOGO. We applied higher order singular value decomposition. One defines the significance of each subtensor in terms of the fraction of the overall information in the data tensor that it captures (Kilmer and Martin, 2004, Ponnapalli et al., 2011). It is proposed that these significant subtensors represent independent biological programs like those previously presented in chapter 3 based on single experiments and eigenanalysis. Similar to classic decomposition, an angular distance of $\pi/4$ indicates a trend that is exclusive to either condition, whereas an angular distance of zero indicates a trend that is common to both the mutant and wildtype datasets. These eigentrends are arranged in decreasing order of their angular distances. Surely, this sequence-independent workflow for more than two experiments, where the variables and operations represent biological reality, will prove incredibly useful in further research studies.

Control theory can be used to steer engineered and natural systems towards a desired state. The goal of such work is to understand how alterations to the circadian network cause changes in the functioning of the circadian clock. This should ideally be asnwered through a mixture of experimental and mathematical analyses of time series data involving measurements of changes in transcript abundance using RNAseq. Here it was attempted using in silico analysis. More insights are being made these days in that growing domain of research. The interaction of distinct units in the circadian system naturally gives rise to complex network structure. This specific circadian system has constantly been in the focus of research for the past decade, with considerable advances in the description of their structural and dynamical properties. However, much less effort has been devoted to studying the controllability of the dynamics taking place within. In the present work we implemented such a methodology to the study of dynamical process of the clock, method initially described by Nepusz and Vicsek based on the edges of a network. They demonstrate that the controllability properties of this process significantly differ from simple nodal dynamics yet these were the part of the analysis too. We demonstrate that transcriptional regulatory networks are particularly easy to control as they fulfill the controllability criteria (Appendix). Analytic calculations show that networks with scale-free degree distributions have better controllability properties than uncorrelated networks, and positively correlated in and out degrees enhance the controllability of the proposed dynamics (Nepusz et al., 2012). As demonstrated in previous chapters that is the case here. It has been shown that one can take control over a relatively low fraction of nodes, but these nodes are incident on more than 90% of the edges in the network and most of these edges have to be controlled explicitly (of course would have to be proved in vivo for a definitive answer). In essence, a dynamical system is controllable if by imposing appropriate external signals on a subset of its nodes, it can be driven from any initial state to any desired state in finite time. The implemented algorithm searches for driver nodes in complex networks with the aim of making them controllable. It operates on both the controllability model of Liu and the switchboard dynamics model of Nepusz and Vicsek (Nepusz et al., 2012). It was no surprise that the driver nodes in order obtained from the aforementioned analysis included LHY followed by the transcription factor, ASG4, and HFR1, a transcription factor which encodes a light-inducible, nuclear bHLH protein involved in phytochrome signaling. To exemplify importance of HFR1 as a hub the mutant phenotype can be considered. The *hfr1* mutants exhibit a long-hypocotyl phenotype only under far-red light but not under red light. They are defective in other phytochrome A-related responses. Mutants also show blue light response defects. Additionally, *HFR1* interacts with *COP1*, co-localizes to nuclear specks and is ubiquinated by *COP1*. It, furthermore, interacts with *PIF4* (bioGRID confirmed), directly providing a link for alteration (providing a link between the light and the clock). *PIF4* has been shown to be a hub for multiple responses to warmer temperature in *Arabidopsis*, including the flowering and hypocotyl elongation. Also among the top list are two other basic helix-loop-helix domain-containing proteins with the entire list. Such data essentially constitutes the first step to the identification of minimum number of driver nodes; ”first step” as the findings would have to be checked against another dataset. It has been shown that switchboard dynamics leads to different controllabilty properties from nodal dynamics and, hence, both were analyzed. Both alternatives are useful here. In terms of the nonlinear dynamics, each node of the system is modeled by a function that performs an operation on the inbound state variables.

The model assumes that the state variables pertain to connected pairs of nodes and that the nodes are not simply passive elements yet active components with information processing capabilities. This has several advantages over the simple nodal dynamics. For example, it is particularly suited in terms of the transcriptional regulatory networks where the mD (fraction of driven edges) score is larger than 0.9, while none of their nD scores (fraction of driver edges) surpass above 0.25 (Nepusz et al., 2012). In other words, one has to take control over a relatively low fraction of nodes, yet these nodes are incident on more than 90% of the edges in the network and the majority of them has to be controlled explicitly, externally. Once the networks have been measured and modeled and decomposed came the time for the deeper understanding of their dynamics and the problem of controllability. As solely few edges are inevitable, that translates into a high degree of robustness against the effects of control. In order to emphasise this point the control path methodology was tested on the gene coexpression data and the results are shown in Fifure 7.1. The driver nodes and control paths are present in Appendix and one such example path together with driver nodes is depicted on Figure 7.1. The data is derived from the coexpression matrix of the At circadian dataset. That quintessentially could be translated into a methodology for the design of mutant lines, with the higher order aim of achieving a desired effect of silencing a pathway for example.

Furthermore, the inference of physical and functional links between cellular components is often reliant upon correlations between experimental measurements, such as gene expression (Barzel and Barabási, 2013a,b). Such correlations are affected by both direct and indirect paths, affecting the ability to identify true pairwise interactions. Barzel and Barabasi proposed a method utilizing the fundamental properties of dynamical correlations in networks to develop a method to silence indirect effects. The method receives as input the observed correlations between node pairs and applies a matrix transformation to turn the correlation matrix into a highly discriminative silenced matrix (terms associated with indirect correlations are silenced). Majority of methods for predicting links assume that the magnitude of Gij correlates with the likelihood of a direct functional/physical link between nodes i and j. Yet Gij cannot distinguish between direct and indirect relationships: a path $i \rightarrow k \rightarrow j$ can result in a measurable response observed between i and j, essentially falsely suggesting the existence of a direct link between them (Barzel and Barabási, 2013a). This enhances only the terms associated with direct causal links. That has been taken advantage of and reduced network was created. The method outperforms many methods, by silencing the indirect correlations by a factor of three for Pearson and Spearman correlations and by a smaller factor of around 2 for MI. This in turn overcomes one of the most common problem of inference methods, the triadic motifs which are three indices subgraphs (Barzel and Barabási, 2013a,b).

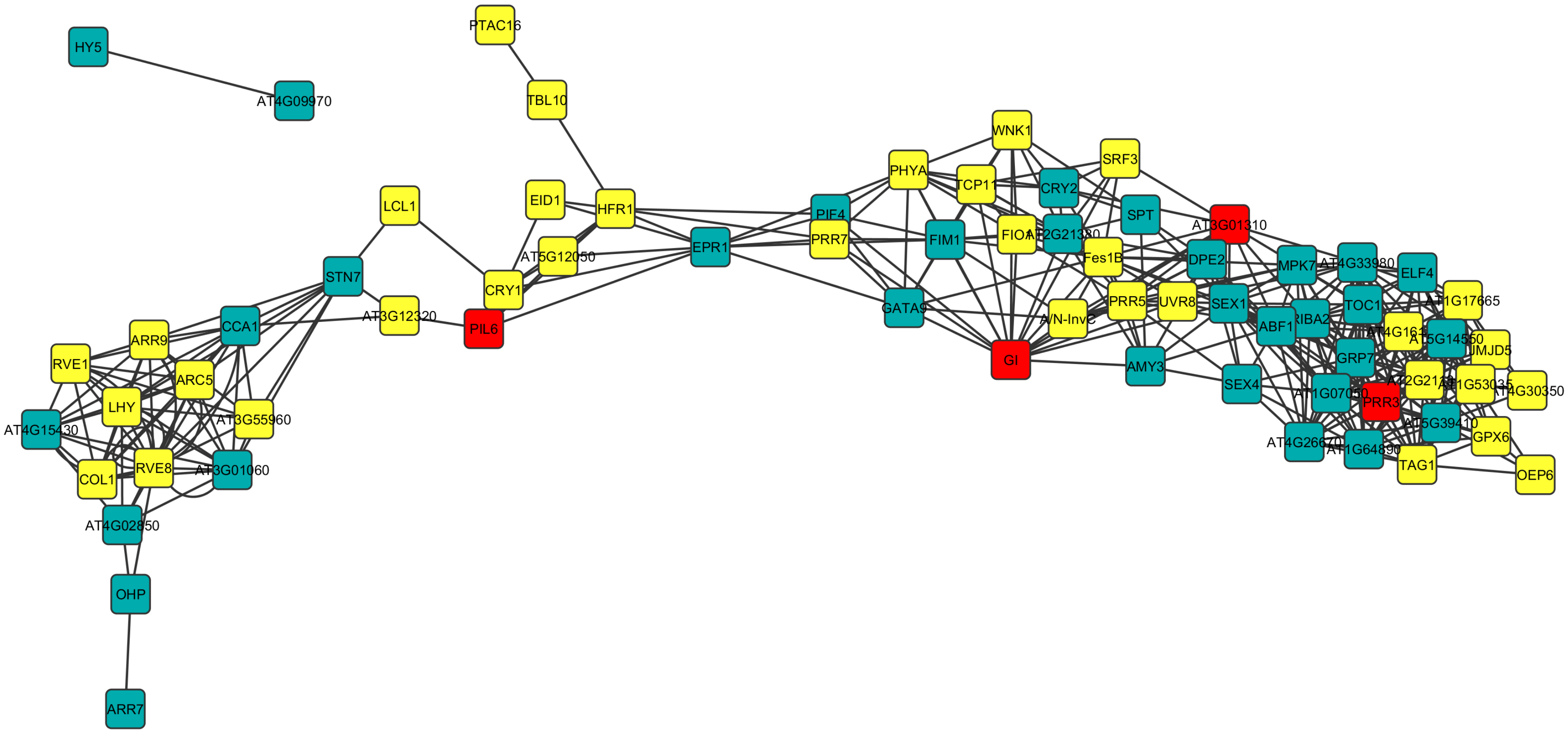


FIGURE 7.1: Circadian control network based on AT coexpression data. The driver nodes in the network are depicted in yellow and the control path is depicted in red. The control path which constitutes central edges to the preservation of robustness of the network includes links between PIL6, GI, PPR3 and phosphoglycerate mutase-like protein (AT3G01310). All the remaining genes are in blue.

Such a applicable approach is avoiding false positives making stringent and conservative inferences and might be worth further analysis in a complex circadian network.

Methods for identification and removal of erroneous links are helpful at cleaning up the noise. The recently proposed methods by Feizi called 'network deconvolution' portrays the measured correlations as a consequence of flows along the edges in the true network. The method proposed by Barzel called 'network silencing' treats the measured correlations as small perturbations that result from adding up the small perturbations induced along edges in the true network. Such inferences may be applied to map the regulatory effect such as for example the direct binding of a TF to the promoter of a target gene. For example the effects of gene A and gene D can be decomposed to the sum of 3 terms: the direct effect of gene B on gene D x by the total effect of gene A on gene B, secondly the direct effect of gene C on gene D multiplied by the total effect of gene A on gene C and thirdly the direct effect of gene F on gene D multiplied by the total effect of gene A on gene F. Given a direct link between genes A and D, that would be seen in the sum. The methods differ in terms of scaling of the links. With regards to partial correlation, it scales each link according to its source and target. This could be applied as alternative for TF:TG regulation as described in chapter 6. The silencing method on the other hand scales the strength of each link according to its target, whereas the method of network deconvolution does not scale the links. Essentially, the strength of each link is used as a proxy for the significance of the association. To put things in perspective, the scaling factors for partial correlation which is the correlation between the residual errors for two variables when they are linearly predicted from all other variables are obtained from the inverse of the correlation matrix.

In this sense, systems biology goes beyond a strict reductionist paradigm, in which the characteristics of the system components are considered in isolation. A key prerequisite for systems methodology is the ability to assay, over time, the state of as many network components as possible. Given microarray data being now supplemented by RNAseq expression data, statistical and mathematical analysis can be used with a great significance. RNAseq provides quality reads for use in genomics studies such as quantification of gene expressions and isoform transcripts identification in a manner superior than conventional methods. It is already starting to benefit crop systems biology will play a crucial role in the understanding of complex crop phenotypes and subsequently crop improvement. This work shows how understanding of the *cis* elements, the organization of modules, the hub genes, their preservation account for complex networks to be rearranged. Whether one considers systems biology to be a paradigm shift or revolution, it will move experimental approaches from a traditional reductionist approach to more holistic treatment of complex biology and that is supported by the present findings. Figure 7.2 depicts for example a low correlation between the microarray probes and

RNAseq tags on several samples taken from the same specimen. This example simply illustrates the potential that RNAseq networks and data might carry in terms of the understnading of the plant clock expression and splicing which likely does not go hand in hand.

7.1.3 FUTURE WORK

The *in silico* work raises the grand question of whether the top down approaches of systems modeling identify a broad solution to the problem of modifying the clock in terms of greater yields and adaptability? Can the omic concepts of systems biology identify the details of the solution at the molecular level? Given that there are n interacting elements, there are $n(n-1)$ possible interactions. If information literally flows in one direction from one element to another in a simple system containing four elements for example, twelve routes exist to carry the information necessary to coordinate the activities of the elements within the system. It is therefore inevitable to supplement these top down approaches with information acquired through the bottom up methodologies.

There are several perspectives from which circadian networks are to be looked upon and investigated. Apart from understanding the topology per se and its dynamics, it is interesting to compare species, the evolution of such networks and spatial differences. All these areas were rather minimally investigated so far. For example the spatial segregation of *GI*, a critical component of plant circadian systems, into nuclear and cytosolic compartments gives rise to differential functions as positive and negative regulators of the circadian core gene, *LHY*, forming an incoherent feed forward loop to regulate *LHY* (Kim et al., 2013). Such results show that spatial and functional segregation of a single molecule species into different cellular compartments gives a way for extending the regulatory capabilities of biological networks. That pertains to differential regulation across tissues with a degree of crosstalk to sustain the robustness of the system. How do tissue-specific clocks exchange temporal information within multicellular organisms? Speaking of applicability of networks in other plant species is another grand challenge yet it is a matter of time. Work in yeast has shown that these combinatorial networks are highly evolvable (Liu et al., 2013). It seems that transcription regulators regulated by multiple factors evolve faster. Furthermore the relationship with the environment causes the type of genes and their optimal time of expression to vary among organisms living under different conditions. Yet, interestingly, it has been shown that natural selection may act upon the output pathways rather than directly on input pathways or on the central oscillator. The supporting evidence includes the notion that transcriptional regulators and their target genes appear to co-evolve as modules.

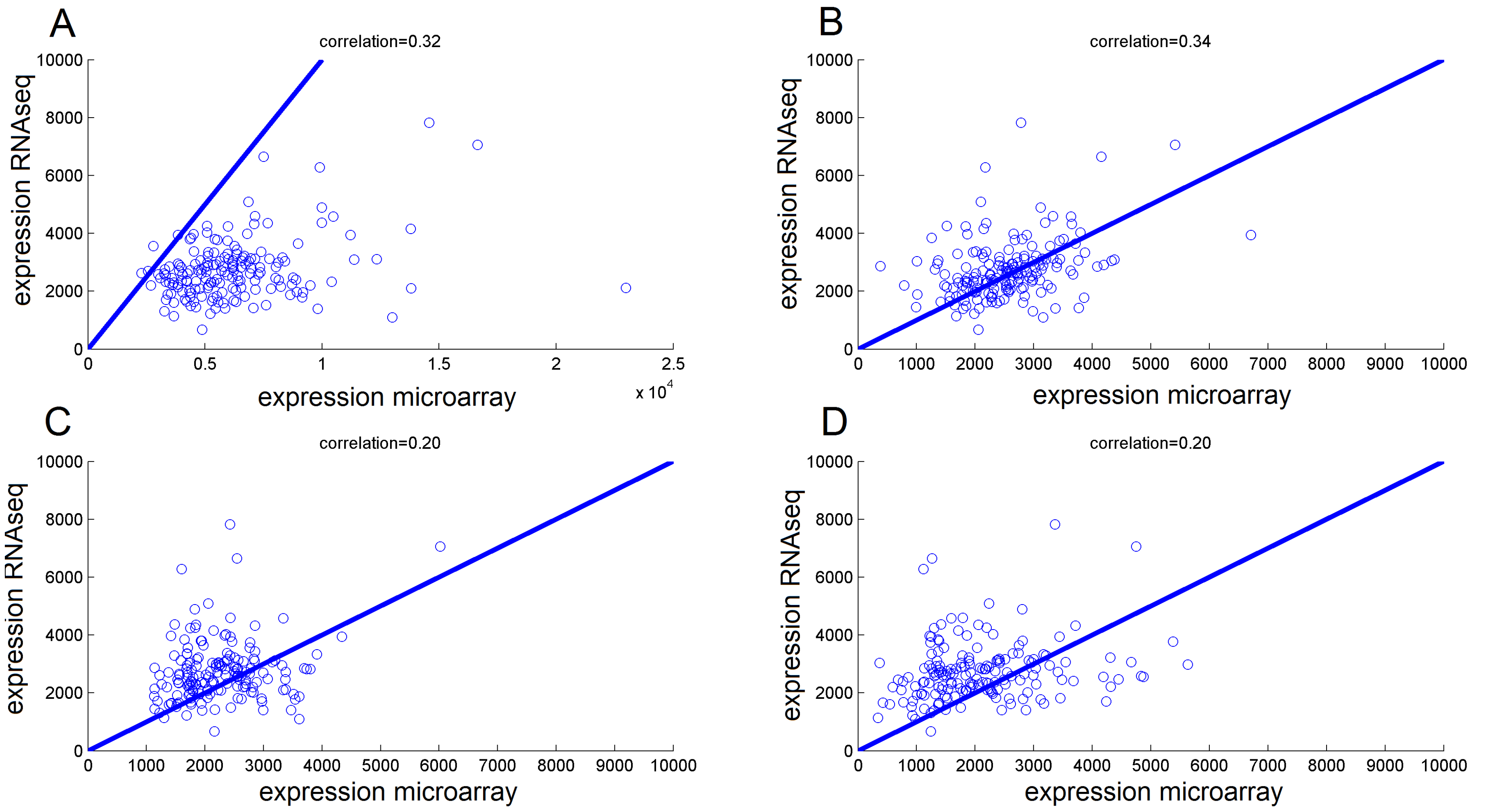


FIGURE 7.2: Figure depicts a low correlation between the microarray probes and RNAseq tags on several samples taken from the same specimen. Multiple dots represent single gene expression of a sample between microarray expression data probes and RNAseq expression tags.

The most agronomically important traits, such as yield and drought tolerance, involve multiple genes and complex interactions with the environment, requiring more sophisticated breeding strategies such as genomic selection. For example, through the inhibition of photosynthesis, it was demonstrated that endogenous oscillations in sugar levels give rise to metabolic feedback to the circadian oscillator through the morning-expressed gene PRR7 (the core eigenmodule gene) (Haydon et al., 2013). As a proof of concept the *prr7* mutants are insensitive to the effects of sucrose on the circadian period. Understanding a system like the plant circadian clock which not solely links the aforementioned processes yet additionally allocates timing to specific output slots is the focal point in achieving the mentioned aims. Combining multiplexing at the sequence level with parallelized sample processing provides biologists with system wide functional testing approaches with sufficient power to match the large scale hypothesis generation that typically results from 'omic' data. Genomewide association studies are an example of the application of genomics in both model plant species and crops that have helped to identify loci and alleles associated with complex traits. The development of high-density molecular markers will be one of the most important tools for informing the design of the breeding programs over the coming years. Speaking strictly of crops, the possibility of generating double haploids to obtain fully homozygous individuals, can significantly speed up breeding, giving rise to opportunities to go faster 'from the genome to the field'. One problem is the delivery method and the second are the intricate links on the plant networks that can potnetially lead to improvement. One such application of direct agronomic relevance is presented in the study of the effects of the circadian clock on postharvest tissues both in terms of chemical content and resistance. It was recently shown that there are links between postharvest entrainment and enhanced herbivore resistance that are present among diverse crops. The differential herbivore resistance of the cabbage tissues after in- versus out-of-phase entrainment provides strong evidence for the successful re-entrainment of the circadian clock in postharvest cabbage tissue (Goodspeed et al., 2013). These results demonstrated that multiple crops subjected to light-dark cycles in phase with *Trichoplusia ni larvae* (T.ni) had lower tissue damage and significantly more tissue remaining after the herbivore challenge (than comparable crops entrained out of phase with T. ni). The future studies will show to what extent the altered phytochemical content, accumulation of glucosinolates, herbivore resistance, and other components of human nutritional value can be incorporated into crops through clock alteration. The optimization of crops through the focus on the central mechanism (clock) linking the core output pathway is by far one of the most exciting areas at the time in terms of crops modification. Considering the balance of experimental work together with the potency of theoretical, the progress curves favor the *in silico* analyses, leaving the experimental work for the most significant verifications. For example, the deep sequencing by RNA-seq of time courses of development can now monitor the dynamic alternative splicing changes in detail integrating these with transcriptional responses. The different phases might use alternative splicing isoforms even though the level of expression does not change (one example could be the alpha and beta isoforms of CCA1). This might not solely be important in the understanding of the physiology but pathophysiology. One other fascinating circadian domain of research that hasn't been exploited enough is a circadian timecourse with regards to plant senesence. As all the cellular functions and fate decisions are governed by spatiotemporal design principles of the circadian clock no doubt these are central to the understanding of the intricate network of dynamic interactions. The individual partners change considerably in time depending on the environment and a multitude of other factors. Inferring the details of this feedback rewiring will allow direct translation to the control of these fractal processes and their spatiotemporal reorganization. Plant fiction is becoming reality. This is shown by studies like that conducted by (Lin, 2014) on increasing yields through utilization of a faster cyanobacterial rubisco rather than their own slower rubisco with higher rates of carbon dioxide fixation per unit of enzyme when compared to control. Another potent recent example is the silencing of a metaphase I specific gene Candidate Pairing homeologous 1 in wheat whose molecular characterization will allow for development of novel alien introgression possibilities (Bhullar et al., 2014). In polyploid wheat, the diploid-like chromosome pairing is under the control of the Ph1 gene through prevention of homeologous chromosome pairing. This candidate Ph1 gene is expressed during meiotic metaphase I and its silencing results in the formation of multivalents like the Ph1 gene mutations. As methods evolve so does the understanding of networks that could be the starting point for modification of this complex system, the plant circadian clock. Given the multiplicity and interactions of the clock with key processes including growth and flowering, the possibilities are endless.

Appendix A

Interview with Steve Kay

How to think big and forge solutions to complex problems On the ticking of the clock and not missing a beat Interview with Professor Steve Kay By Sandra Smieszek

It is certainly my great pleasure to introduce Professor Steve Kay holder of the Anna H. Bing Dean's Chair and professor of biological sciences, leader, educator and innovator. He is a member of the National Academy of Sciences, and a fellow of the American Association for the Advancement of Science. The interview is providing us not solely with the story of his expansive career and additionally shows us how it influenced research and education. Professor Steve Kay is a world expert on circadian rhythms. He spent two decades identifying the photoreceptors, genes, and complex networks that make these internal clocks tick. Steve Kay transformed the field of molecular biology. His famous tricks include the blinking mustard plants and glowing fruit flies to explore the molecular genetic basis of circadian clocks in plants, flies, and mammals. His mantra, as the 21st dean of USC Dornsife: educate, enrich and empower . . . sums it all up.

1. What influences directed you to your specific area of research? Who influenced your scientific thinking early in your career, and how?
   I became interested in biology early in my childhood. It all began on the small island of Jersey, off the coast of Normandy. Many of my family members were fisherman, and I spent a lot of time on commercial boats. These oddities of marine life coupled with teachers and the first microscope predestined my career direction. Certainly my supervisors pushed me to 'think big'. Trevor Griffiths who was my Ph.D. supervisor introduced me to the world of plants. It was during my doctoral studies when I discovered that light regulated the expression of the gene that produced the enzyme for chlorophyll synthesis. These were the beginnings of the day/night cycle observations that set the scene. It was Trevor Griffiths who advised me to pursue my research in United States. That is when I started a

postdoctoral fellowship at a lab of Nam-Hai Chua who focused on light dependent gene expression in plants. He certainly taught me how to approach more than one thing at a time. These were incredibly exciting times when we worked on the first vectors for transgenic plants.

2. What scientific breakthrough over the past couple of years influenced your research directions and why/how?
   The 'eureka' moment, it was definitely during my postdoctoral studies. Light signals change in gene expression patterns, I am thinking here particularly of chlorophyll a/b binding CAB gene. That is when we conducted the experiments around the clock. The discovery essentially showed how CAB was regulated by the circadian clock. That was exactly 1985 and that was the first direct evidence for the role of circadian rhythm exerting its effect at a molecular level. That was astonishing.

3. What was the most difficult stage in Your career?
   Probably I had several, science is really hard . . . , long periods of failure intermitted by splices of success. I can say here that cloning of TOC1 gene took us quite some time, 5 years (published in Science in 1995). Of course that was back in the days. It took many more years to elucidate what the gene does.

4. What recent developments in basic plant science are influencing policy making bodies today?
   I think it varies a lot by region. It seems more difficult to convince that funding research to gain knowledge in reference species is still valid and crucial. Overall less than 1% of general funding goes to plant science versus around 30% in China, 20% Europe. That is entirely different to what it was and is supposed to be. I would like to highlight and call for appreciation of the critical role of robust funding for the basic sciences which, provided will lay the foundation for improvements in health, agriculture and the environment.

5. What advice would you give to a student interested in plant biology today?
   To be concerned wide and go deep

6. As an employer, what are the five key qualities you look for in a potential team member?
   Passion, effort, intellectual capability, discipline, and horsepower

7. "The challenge for biologists is to become comfortable with mathematical tools" could You elaborate?
   Of course, beautiful examples of what can be done specifically in our domain come from Andrew Millar. Nowadays it is instrumental for biologists to become

comfortable with mathematical tools. At the same time we have to be comfortable with biology becoming a predictive science. Bottom up approaches, painstakingly craft models and great emphasis should be placed on these. Whereas top down approaches with the present capabilities should clearly be incorporated, it is not one instead of the other. It seems the present mission of systems biology will be the fusion of both.

8. "I've watched agog as the word MOOC has proliferated and spiraled into the higher education buzzword of the year." Speaking of the new wave of educators what is Your stand on the evolution of coursera?
   It is fantastic yet it will never be a replacement. As useful as it can be, it is superficial at the same moment. I have a direct example coming from John Hogenesch who runs one of these classes (https://www.coursera.org/course/genomescience). The numbers were astonishing, 10,000 people enrolled, 4400 participated, 822 took exam, versus numbers that come to class that range in 20s. Yet looking at the numbers, it seems these are professional, that participate. The opportunity is once indescribable, and after all who hasn't used Kahn academy? It seems like the optimal refreshment.

9. With genomics monopolizing attention what do You think is the next buzz domain that will take over in the years to come?
   I think high throughput sequencing in all shapes and sizes, together with post-translational studies will keep us busy in the upcoming years

10. Reductionism, as a paradigm, is expired, and complexity, as a field, is tired. Data-based mathematical models of complex systems are offering a fresh perspective, rapidly developing into a new discipline: network science. What can network science do for plant biology in reality?
    That is the way to go and I am all for these studies, as long as they complement, research. Moreover there is true potential in the study of dynamics of biological systems like done by Trey Ideker. This, coupled with the wealth of high throughout data is truly exciting.

11. Who should and will fund future molecular biology research, what is the interaction between government funding/private, commercial/charitable donations.
    That truly varies upon regions and projects so it is difficult to elaborate

12. Once speaking of the present stand on sequencing You made a comment "It's comparable to Darwin's theory of evolution," do You agree now?
    Certainly, it is a massive revelation. Nevertheless it is complementary. It is the variation beyond nucleotide that constitutes lots of the present conundrums that one has to focus on

13. On a light note what is Your favorite book?

    'Do Androids Dream of Electric Sheep?' By Philip K. Dick.

Appendix B

Interview with Stephen Altschul

On BLAST, BRCA1 and the Wild Duck Stephen Altschul By Sandra Smieszek Interview with creator of the basic local alignment search tool (BLAST) and a eminent bioinformatics forerunner and literati does not call for much introduction. Stephen Altschul graduated summa cum laude from Harvard University, and has a Ph.D. in the same field of research from MIT. What can BLAST do for us is something we all know yet the quintessence is in how it originated and even more interestingly who is the man behind the scene. It is certainly my great pleasure to introduce Stephen Altschul providing us not solely with the story of his algorithms yet additionally of the power-law explosion of bioinformatics over the past decade.

SS BLAST was published in Journal of Molecular Biology in 1990. Since that time it has been cited over 43,568 times, what a feeling?
Certainly accomplished, it was designed to be faster than FASTA at finding very strong similarities. It was something of a surprise that it performed as well as it did at finding weak similarities as well.
SS What influences directed you to your specific area of research? Who influenced your scientific thinking early in your career, and how?
Having graduated, I spent a lot of time reading on the potential applied mathematical problems in biology. Among the inspirational books I read a textbook entitled "Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence" Comparison by Sankoff and Kruskal. I read the "Double Helix" by James Watson. I traveled a lot to conferences yet speaking of individuals that had particular impact, David J. Lipman, the present director of NCBI was my great inspiration. That is a route I took from mathematics per se to the world bioinformatics.
SS What scientific discovery over the past couple of years had a major impact upon You?

The most exciting discovery I was involved with as it unfolded was certainly the characterization of BRCA1 in early 1990s. It was a perfect example of applying sequence alignment tools for significant discovery of functional motifs of BRCA1. I want to credit Peer Bork along with Eugene Koonin. We mapped out the functional motifs. BRCA1 was partitioned into globular and not globular domains. We have noticed a similarity between the designated 53BP1 that has been identified by its ability to bind p53. The other hits included KIAA0170 and RAD9. "The probability exceeded 87% that a pattern as strong as the previously noted 'granin motif' would be shared by a random sequence as long as BRCA1 and the then-extant motif database, thus lending no statistical support to the relevance of this motif." Now the C-terminus of BRCA1 is known to contain two 95-residue BRCT domains, which are also found in many other proteins involved in DNA repair and cell cycle regulation. The crystal structure was later defined. It is not solely the story of characterization of one of the most important tumor suppressor genes in cancer yet additionally the story of how well applied statistics can shed light into true positive interesting domains in this example.

SS What was the most difficult stage in Your career?

I guess right after graduation getting applied problems was the most difficult stage yet it did not last long. I ended up working in a 'lucky field' one that has grown rapidly over the past decade.

SS Speaking of the new wave of educators what is Your stand on the evolution of coursera, upon MOOC which has spiralled into the higher education buzzword of the year?

Certainly beneficial for the society, even I myself sometimes benefit from the aspects of online lectures and series. Nevertheless long term consequences may be ambiguous.

SS Reductionism, as a paradigm, is expired, and complexity, as a field, is tired. Data-based mathematical models of complex systems are offering a fresh perspective, rapidly developing into a new discipline: network science. Do You subscribe to that view?

It's difficult for me to say 'no' although it is not my field of expertise, it certainly sounds interesting at first glance.

SS Who should and will fund future bioinformatic research, what is the interaction between gov funding/private, commecrcial/charitable donations?

(With laughter) I might not be the correct person to ask as I had was lucky enough not to have applied to grants.

SS Speaking of equal animals, what ongoing ethical dilemmas is the present society facing in the light of present technological advances?

In fact we have a lot of such conversations here at NIH. We are facing a sort of Wild Duck dilemma of whether attaining truth at all costs is the desired destination. It is still to be answered if the truth will out-compete the ruins brought to the 'household' (that is abuse of the system) (for SA I am referring to the Wild Duck and the view of Greger's)

SS You have monitored the rapid growth and expansion of the field, what do You think is the next big route bioinformatics science will take?

Big route for bioinformatics science: The term currently comprehends a vary large range of disciplines, which deal with data as disparate as medical records, literature, sequences, expression patterns, mass spectroscopy, protein interactions, etc. For that bioinformatics which has been focused on research in molecular biology, a challenge will be to create tools that are reliable and scalable enough to be useful in clinical practice.

SS What advice would You give students starting their bioinformatics careers?

Advice to people starting out in bioinformatics: If you comefrom the fields of computer science, math, or physics, learn as much biology as you can. If you come from any field, learn a lot of statistics.

SS If You could provide us with anything that comes to Your mind when You hear the following extremely high throughput:

faster algorithms

tool You are most proud of:

PSI-BLAST, which for the first time made "protein profile" searches accessible to the non-expert

scientific superhero:

Mendel and Turing

Often read:

the newspaper

If you had a billion dollars to fund research or charity, where would it go?

It is not my professional field, but I am most concerned with the degradation of the planet's environment. Many approaches have been tried to address this increasingly urgent problem, but ground continues to be lost.

THANK YOU

Appendix C

Pseudocode

All remaining supplementary material attached on external media.

Bibliography

Alabadí, D., Oyama, T., Yanovsky, M.J., Harmon, F.G., Más, P., Kay, S.A.: Reciprocal regulation between TOC1 and LHY/CCA1 within the Arabidopsis circadian clock. Science 293(5531), 880–3 (2001)

Albert, R.: Boolean Modelingof Genetic Regulatory Networks. In E. Ben-Naim, H. Frauenfelder, Z. Toroczkai, editors, Complex Networks, volume 650 of Lecture Notes in Physics, 459–481. Springer Berlin Heidelberg (2004)

Albert, R.: Scale-free networks in cell biology. Journal of Cell Science 118(21), 4947–4957 (2005)

Albert, R.: Network inference, analysis, and modeling in systems biology. The Plant cell 19(11), 3327–38 (2007)

Allen, J.E., Majoros, W.H., Pertea, M., Salzberg, S.L., et al.: JIGSAW, GeneZilla, and GlimmerHMM: puzzling out the features of human genes in the ENCODE regions. Genome Biol 7(Suppl 1), 1–13 (2006)

Alon, U.: Biological networks: the tinkerer as an engineer. Science 301(5641), 1866–7 (2003)

Alter, O.: Singular value decomposition for genome-wide expression data processing and modeling. Proceedings of the National Academy of Sciences 97(18), 10101–10106 (2000)

Alter, O., Brown, P.O., Botstein, D.: Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. Proceedings of the National Academy of Sciences of the United States of America 100(6), 3351–6 (2003)

Ashkenazy, H., Erez, E., Martz, E., Pupko, T., Ben-Tal, N.: ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic acids research 38(suppl 2), W529–W533 (2010)

Bader, G.D., Betel, D., Hogue, C.W.V.: BIND: the Biomolecular Interaction Network Database. Nucleic acids research 31(1), 248–50 (2003)

Bailey, T.L., Williams, N., Misleh, C., Li, W.W.: MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic acids research 34(suppl 2), W369–W373 (2006)

Bandyopadhyay, S., Mehta, M., Kuo, D., Sung, M.K., Chuang, R., Jaehnig, E.J., Bodenmiller, B., Licon, K., Copeland, W., Shales, M., Fiedler, D., Dutkowski, J., Guénolé, A., van Attikum, H., Shokat, K.M., Kolodner, R.D., Huh, W.K., Aebersold, R., Keogh, M.C., Krogan, N.J., Ideker, T.: Rewiring of genetic networks in response to DNA damage. Science 330(6009), 1385–9 (2010)

Barabási, A.L.: The network takeover. Nature Physics 8(1), 14–16 (2011)

Barabási, A.L.: Network science. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 371(1987) (2013)

Barabási, A.L., Oltvai, Z.N.: Network biology: understanding the cell's functional organization. Nature reviews. Genetics 5(2), 101–13 (2004)

Barzel, B., Barabási, A.L.: Network link prediction by global silencing of indirect correlations. Nature biotechnology 31(8), 720–5 (2013a)

Barzel, B., Barabási, A.L.: Universality in network dynamics. Nature Physics 9(10), 673–681 (2013b)

Bean, G.J., Ideker, T.: Differential analysis of high-throughput quantitative genetic interaction data. Genome biology 13(12), R123 (2012)

Beer, M.a., Tavazoie, S.: Predicting gene expression from sequence. Cell 117(2), 185–98 (2004)

Belhaj, K., Chaparro-Garcia, A., Kamoun, S., Nekrasov, V.: Plant genome editing made easy: targeted mutagenesis in model and crop plants using the CRISPR/Cas system. Plant Methods 9(1), 39 (2013)

Beltrami, E.: Sulle funzioni bilineari. Giornale di Matematiche as Uso degli Studenti Della Universita 11, 98–106 (1873)

Bhullar, R., Nagarajan, R., Bennypaul, H., Sidhu, G.K., Sidhu, G., Rustgi, S., von Wettstein, D., Gill, K.S.: Silencing of a metaphase I-specific gene results in a phenotype similar to that of the Pairing homeologous 1 (Ph1) gene mutations. Proceedings of the National Academy of Sciences 111(39), 14187–14192 (2014)

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.: Complex networks: Structure and dynamics. Physics Reports 424(4-5), 175–308 (2006)

Breiman, L.: Random forests. Machine learning 45(1), 5–32 (2001)

Bujdoso, N., Davis, S.J.: Mathematical modeling of an oscillating gene circuit to unravel the circadian clock network of Arabidopsis thaliana. Frontiers in Plant Sciences 4(3), 1–8 (2013)

Bussemaker, H.J., Li, H., Siggia, E.D.: Regulatory element detection using correlation with expression. Nature genetics 27(2), 167–71 (2001)

Carré, I., Veflingstad, S.R.: Emerging design principles in the Arabidopsis circadian clock. Seminars in cell & developmental biology 24(5), 393–8 (2013)

Chang, X., Xu, T., Li, Y., Wang, K.: Dynamic modular architecture of protein-protein interaction networks beyond the dichotomy of 'date' and 'party' hubs. Scientific reports 3, 1691 (2013)

Chow, B.Y., Kay, S.a.: Global approaches for telling time: Omics and the Arabidopsis circadian clock. Seminars in cell & developmental biology 24(5), 383–92 (2013)

Cope, L.: Statistical Properties of the Integrative Correlation Coefficient : a Measure of Cross-study Gene Reproducibility. Johns Hopkins University, Department of Biostatistics Working Papers (2011)

Covington, M.F., Maloof, J.N., Straume, M., Kay, S.A., Harmer, S.L.: Global transcriptome analysis reveals circadian regulation of key pathways in plant growth and development. Genome biology 9(8), R130 (2008)

Cox, R.S., Nishikata, K., Shimoyama, S., Yoshida, Y., Matsui, M., Makita, Y., Toyoda, T.: PromoterCAD: Data-driven design of plant regulatory DNA. Nucleic acids research 41(Web Server issue), W569–74 (2013)

Devlin, P.F.: Signs of the time: environmental input to the circadian clock. Journal of Experimental Botany 53(374), 1535–1550 (2002)

Devlin, P.F., Kay, S.A.: Cryptochromes are required for phytochrome signaling to the circadian clock but not for rhythmicity. The Plant cell 12(12), 2499–2510 (2000)

Dodd, A.N., Salathia, N., Hall, A., Kévei, E., Tóth, R., Nagy, F., Hibberd, J.M., Millar, A.J., Webb, A.A.R.: Plant circadian clocks increase photosynthesis, growth, survival, and competitive advantage. Science 309(5734), 630–3 (2005)

Doherty, C.J., Kay, S.A.: Circadian control of global gene expression patterns. Annual review of genetics 44, 419–444 (2010)

Dondelinger, F., Husmeier, D., Lèbre, S.: Dynamic Bayesian networks in molecular plant science: inferring gene regulatory networks from multiple gene expression time series. Euphytica 183(3), 361–377 (2012)

Dong, J., Horvath, S.: Understanding network concepts in modules. BMC systems biology 1(1), 24 (2007)

Doyle, M.R., Davis, S.J., Bastow, R.M., McWatters, H.G., Kozma-Bognár, L., Nagy, F., Millar, A.J., Amasino, R.M.: The ELF4 gene controls circadian rhythms and flowering time in Arabidopsis thaliana. Nature 419(6902), 74–7 (2002)

Dutkowski, J., Kramer, M., Surma, M.A., Balakrishnan, R., Cherry, J.M., Krogan, N.J., Ideker, T.: A n A ly s i s A gene ontology inferred from molecular networks. Nature 31(1) (2013)

Eckart, C., Young, G.: The approximation of one matrix by another of lower rank. Psychometrika 1, 211–218 (1936)

Edgar, R.S., Green, E.W., Zhao, Y., van Ooijen, G., Olmedo, M., Qin, X., Xu, Y., Pan, M., Valekunja, U.K., Feeney, K.A., Maywood, E.S., Hastings, M.H., Baliga, N.S., Merrow, M., Millar, A.J., Johnson, C.H., Kyriacou, C.P., O'Neill, J.S., Reddy, A.B.: Peroxiredoxins are conserved markers of circadian rhythms. Nature 485(7399), 459–64 (2012)

Edwards, K.D., Akman, O.E., Knox, K., Lumsden, P.J., Thomson, A.W., Brown, P.E., Pokhilko, A., Kozma-Bognar, L., Nagy, F., Rand, D.A., Millar, A.J.: Quantitative analysis of regulatory flexibility under changing environmental conditions. Molecular Systems Biology 6(424), 424 (2010)

Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences 95(25), 14863–14868 (1998)

Elowitz, M.B., Leibler, S.: A synthetic oscillatory network of transcriptional regulators. Nature 403(6767), 335–8 (2000)

Emmert-Streib, F., Dehmer, M., Haibe-Kains, B.: Gene regulatory networks and their applications: Understanding biological and medical problems in terms of networks. Frontiers in Cell and Developmental Biology 2(38) (2014)

Endo, M.: Tissue-specific clocks in Arabidopsis show asymmetric coupling. Nature 515(7527) (2014)

Espinoza, C., Bieniawska, Z., Hincha, D.K., Hannah, M.A.: Interactions between the circadian clock and cold-response in Arabidopsis. Plant Signaling and Behavior 3(8), 593–594 (2008)

Esvelt, K.M., Wang, H.H.: Genome-scale engineering for systems and synthetic biology. Molecular systems biology 9(641), 641 (2013)

Feiglin, A., Hacohen, A., Sarusi, A., Fisher, J., Unger, R., Ofran, Y.: Static network structure can be used to model the phenotypic effects of perturbations in regulatory networks. Bioinformatics 28(21), 2811–8 (2012)

Filichkin, S.A., Priest, H.D., Givan, S.A., Shen, R., Bryant, D.W., Fox, S.E., Wong, W.K., Mockler, T.C.: Genome-wide mapping of alternative splicing in Arabidopsis thaliana. Genome research 20(1), 45–58 (2010)

Fortunato, S.: Community detection in graphs. Physics Reports 486(3-5), 75–174 (2010)

Friedman, J.H.: On bias, variance, 0/1—loss, and the curse-of-dimensionality. Data mining and knowledge discovery 1(1), 55–77 (1997)

Friedman, N.: Using Bayesian networks to analyze expression data. J Comput Biol. 7, 3–4 (2000)

Fujiwara, S., Wang, L., Han, L., Suh, S.S., Salomé, P.A., McClung, C.R., Somers, D.E.: Post-translational regulation of the Arabidopsis circadian clock through selective proteolysis and phosphorylation of pseudo-response regulator proteins. The Journal of biological chemistry 283(34), 23073–83 (2008)

Gardner, T.S., Cantor, C.R., Collins, J.J.: Construction of a genetic toggle switch in Escherichia coli. Nature 403(6767), 339–42 (2000)

Gardner, T.S., Faith, J.J.: Reverse-engineering transcription control networks. Physics of life reviews 2(1), 65–88 (2005)

Gendron, J.M., Pruneda-Paz, J.L., Doherty, C.J., Gross, A.M., Kang, S.E., Kay, S.A.: Arabidopsis circadian clock protein, TOC1, is a DNA-binding transcription factor. Proceedings of the National Academy of Sciences 109(8), 3167–3172 (2012)

Ghoshal, G., Barabási, A.L.: Ranking stability and super-stable nodes in complex networks. Nature communications 2, 394 (2011)

Giuliano, G., Hoffman, N.E., Ko, K., Scolnik, P.A., Cashmore, A.R.: A light-entrained circadian clock controls transcription of several plant genes. The EMBO journal 7(12), 3635–42 (1988)

Gjuvsland, A., Plahte, E., Omholt, S.: Threshold-dominated regulation hides genetic variation in gene expression networks. BMC Systems Biology 1(1), 57 (2007)

Goff, S.A., Vaughn, M., McKay, S., Lyons, E., Stapleton, A.E., Gessler, D., Matasci, N., Wang, L., Hanlon, M., Lenards, A., et al.: The iPlant collaborative: cyberinfrastructure for plant biology. Frontiers in plant science 2 (2011)

Goldstein, B.a., Polley, E.C., Briggs, F.B.S.: Random forests for genetic association studies. Statistical applications in genetics and molecular biology 10(1), 32 (2011)

Goodspeed, D., Liu, J.D., Chehab, E.W., Sheng, Z., Francisco, M., Kliebenstein, D.J., Braam, J.: Postharvest circadian entrainment enhances crop pest resistance and phytochemical cycling. Current biology 23(13), 1235–41 (2013)

Graf, A., Schlereth, A., Stitt, M., Smith, A.M.: Circadian control of carbohydrate availability for growth in Arabidopsis plants at night. Proceedings of the National Academy of Sciences 107(20), 9458–9463 (2010)

Granger, C.W.: Investigating causal relations by econometric models and cross-spectral methods. Econometrica: Journal of the Econometric Society 424–438 (1969)

Günther, B., Morgado, E.: Dimensional analysis revisited. Biological research 36(3-4), 405–10 (2003)

Han, J.D.J., Bertin, N., Hao, T., Goldberg, D.S., Berriz, G.F., Zhang, L.V., Dupuy, D., Walhout, A.J.M., Cusick, M.E., Roth, F.P., Vidal, M.: Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature 430(6995), 88–93 (2004)

Hanano, S., Domagalska, M.A., Nagy, F., Davis, S.J.: Multiple phytohormones influence distinct parameters of the plant circadian clock. Genes to Cells 11(12), 1381–1392 (2006)

Harmer, S.L.: Orchestrated Transcription of Key Pathways in Arabidopsis by the Circadian Clock. Science 290(5499), 2110–2113 (2000)

Hartwell, L.H., Hopfield, J.J., Leibler, S., Murray, A.W.: From molecular to modular cell biology. Nature 402(6761 Suppl), C47–52 (1999)

Haydon, M.J., Mielczarek, O., Robertson, F.C., Hubbard, K.E., Webb, A.A.R.: Photosynthetic entrainment of the Arabidopsis thaliana circadian clock. Nature 502(7473), 689–92 (2013)

Hayes, K.R., Beatty, M., Meng, X., Simmons, C.R., Habben, J.E., Danilevskaya, O.N.: Maize global transcriptomics reveals pervasive leaf diurnal rhythms but rhythms in developing ears are largely limited to the core oscillator. PloS one 5(9), e12887 (2010)

Hempel, S., Koseska, A., Kurths, J., Nikoloski, Z.: Inner composition alignment for inferring directed networks from short time series. Physical review letters 107(5), 054101 (2011)

Holter, N.S., Maritan, A., Cieplak, M., Fedoroff, N.V., Banavar, J.R.: Dynamic modeling of gene expression data. Proceedings of the National Academy of Sciences 98(4) (2001)

Hsu, P.Y., Devisetty, U.K., Harmer, S.L.: Accurate timekeeping is controlled by a cycling activator in Arabidopsis. eLife 2, e00473 (2013)

Hu, Y., Hu, S., Wang, W., Wu, X., Marshall, F.B., Chen, X., Hou, L., Wang, C.: Earliest evidence for commensal processes of cat domestication. Proceedings of the National Academy of Sciences 111(1), 116–120 (2014)

Huang, D.W., Sherman, B.T., Lempicki, R.a.: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols 4(1), 44–57 (2009)

Huang, W., Pérez-García, P., Pokhilko, A., Millar, A.J., Antoshechkin, I., Riechmann, J.L., Mas, P.: Mapping the core of the Arabidopsis circadian clock defines the network structure of the oscillator. Science 336(6077), 75–9 (2012)

Hughes, J.D., Estep, P.W., Tavazoie, S., Church, G.M.: Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. Journal of molecular biology 296(5), 1205–14 (2000)

Huynh-Thu, V.A., Irrthum, A., Wehenkel, L., Geurts, P.: Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. PLoS ONE 5, e12776 (2010)

Iasonos, A., Schrag, D., Raj, G.V., Panageas, K.S.: How to build and interpret a nomogram for cancer prognosis. Journal of Clinical Oncology 26(8), 1364–1370 (2008)

Ideker, T., Krogan, N.J.: Differential network biology. Molecular systems biology 8, 565 (2012)

Ingkasuwan, P., Netrphan, S., Prasitwattanaseree, S., Tanticharoen, M., Bhumiratana, S., Meechai, A., Chaijaruwanich, J., Takahashi, H., Cheevadhanarak, S.: Inferring transcriptional gene regulation network of starch metabolism in Arabidopsis thaliana leaves using graphical Gaussian model. BMC Systems Biology 6(1), 100 (2012)

Ishwaran, H.: The effect of splitting on random forests. Machine Learning 1–44 (2014)

Joon, S.P.: A Self-Regulatory Circuit of CIRCADIAN CLOCK-ASSOCIATED1 Underlies the Circadian Clock Regulation of Temperature Responses in Arabidopsis. Plant Cell 111(1), 116–120 (2012)

Khan, S.: Coordination of the maize transcriptome by a conserved circadian clock. BMC Plant Biol 10(126) (2010)

Kholodenko, B., Yaffe, M.B., Kolch, W.: Computational approaches for analyzing information flow in biological networks. Science Signaling 5(220), re1 (2012)

Kholodenko, B.N., Kiyatkin, A., Bruggeman, F.J., Sontag, E., Westerhoff, H.V., Hoek, J.B.: Untangling the wires: a strategy to trace functional interactions in signaling and gene networks. Proceedings of the National Academy of Sciences 99(20), 12841–12846 (2002)

Kilmer, M.E., Martin, C.D.M.: Decomposing a tensor. SIAM News 37(9), 19–20 (2004)

Kim, Y., Han, S., Yeom, M., Kim, H., Lim, J., Cha, J.Y., Kim, W.Y., Somers, D.E., Putterill, J., Nam, H.G., Hwang, D.: Balanced nucleocytosolic partitioning defines a spatial network to coordinate circadian physiology in plants. Developmental cell 26(1), 73–85 (2013)

Kitano, H.: Biological robustness. Nature Reviews Genetics 5(11), 826–837 (2004)

Kloppstech, K.: Diurnal and circadian rhythmicity in the expression of light-induced plant nuclear messenger RNAs. Planta 165(4), 502–506 (1985)

Korenčič, A., Bordyugov, G., Rozman, D., Goličnik, M., Herzel, H., et al.: The Interplay of cis-Regulatory Elements Rules Circadian Rhythms in Mouse Liver. PloS one 7(11), e46835 (2012)

Krishna, R., Li, C.T., Buchanan-Wollaston, V.: A temporal precedence based clustering method for gene expression microarray data. BMC Bioinformatics 11(1), 68 (2010)

Kuno, N., Møller, S.G., Shinomura, T., Xu, X., Chua, N.H., Furuya, M.: The novel MYB protein EARLY-PHYTOCHROME-RESPONSIVE1 is a component of a slave circadian oscillator in Arabidopsis. The Plant Cell Online 15(10), 2476–2488 (2003)

Langfelder, P., Luo, R., Oldham, M.C., Horvath, S.: Is My Network Module Preserved and Reproducible? PLoS Computational Biology 7(1), e1001057 (2011)

Lee, J.W., Lee, J.B., Park, M., Song, S.H.: An extensive comparison of recent classification tools applied to microarray data. Computational Statistics & Data Analysis 48(4), 869–885 (2005)

Li, G., Siddiqui, H., Teng, Y., Lin, R., Wan, X.Y., Li, J., Lau, O.S., Ouyang, X., Dai, M., Wan, J., Devlin, P.F., Deng, X.W., Wang, H.: Coordinated transcriptional regulation underlying the circadian clock in Arabidopsis. Nature cell biology 13(5), 1–9 (2011a)

Li, G., Siddiqui, H., Teng, Y., Lin, R., Wan, X.y., Li, J., Lau, O.S., Ouyang, X., Dai, M., Wan, J., Devlin, P.F., Deng, X.W., Wang, H.: Coordinated transcriptional regulation underlying the circadian clock in Arabidopsis. Nature cell biology 13(5), 616–22 (2011b)

Liebermeister, W.: Linear modes of gene expression determined by independent component analysis. Bioinformatics 18(1), 51–60 (2002)

Lin, M.: A faster Rubisco with potential to increase photosynthesis in crops. Nature 513(7519) (2014)

Lin, R., Ding, L., Casola, C., Ripoll, D.R., Feschotte, C., Wang, H.: Transposase-derived transcription factors regulate light signaling in Arabidopsis. Science 318(5854), 1302–5 (2007)

Liu, Y.Y., Slotine, J.J., Barabási, A.L.: Controllability of complex networks. Nature 473(7346), 167–173 (2011)

Liu, Y.Y., Slotine, J.J., Barabási, A.L.: Observability of complex systems. Proceedings of the National Academy of Sciences 110(7), 2460–2465 (2013)

Lopes, M.: Experimental assessment of static and dynamic algorithms for gene regulation inference from time series expression data. Frontiers in Genetics 4, 3–4 (2013)

López-Kleine, L., Leal, L., López, C.: Biostatistical approaches for the reconstruction of gene co-expression networks based on transcriptomic data. Briefings in functional genomics 12(5), 457–67 (2013)

Lorenz, D.M., Jeng, A., Deem, M.W.: The emergence of modularity in biological systems. Physics of life reviews 8(2), 129–160 (2011)

Lunetta, K.L., Hayward, L.B., Segal, J., Van Eerdewegh, P.: Screening large-scale association study data: exploiting interactions using random forests. BMC genetics 5(1), 32 (2004)

Marbach, D., Prill, R.J., Schaffter, T., Mattiussi, C., Floreano, D., Stolovitzky, G.: Revealing strengths and weaknesses of methods for gene network inference. Proceedings of the National Academy of Sciences 107(14), 6286–91 (2010)

Margolin, A.A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., Califano, A.: ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 7(Suppl 1), S7 (2006)

Markus, B.: Evening Expression of ArabidopsisGIGANTEA Is Controlled by Combinatorial Interactions among Evolutionarily Conserved Regulatory Motifs. Plant Cell 111(1), 116–120 (2014)

Maxwell, B.B., Andersson, C.R., Poole, D.S., Kay, S.A., Chory, J.: HY5, Circadian Clock-Associated 1, and a cis-element, DET1 dark response element, mediate DET1 regulation of chlorophyll a/b-binding protein 2 expression. Plant physiology 133(4), 1565–77 (2003)

McWatters, H.G., Devlin, P.F.: Timing in plants–a rhythmic arrangement. FEBS letters 585(10), 1474–1484 (2011)

McWatters, H.G., Kolmos, E., Hall, A., Doyle, M.R., Amasino, R.M., Gyula, P., Nagy, F., Millar, A.J., Davis, S.J.: ELF4 is required for oscillatory properties of the circadian clock. Plant physiology 144(1), 391–401 (2007)

Michael, T.P., McClung, C.R.: Phase-specific circadian clock regulatory elements in Arabidopsis. Plant Physiology 130(2), 627–638 (2002)

Michael, T.P., Mockler, T.C., Breton, G., McEntee, C., Byer, A., Trout, J.D., Hazen, S.P., Shen, R., Priest, H.D., Sullivan, C.M., Givan, S.a., Yanovsky, M., Hong, F., Kay, S.a., Chory, J.: Network discovery pipeline elucidates conserved time-of-day-specific cis-regulatory modules. PLoS genetics 4(2), e14 (2008)

Millar, A., Carre, I., Strayer, C., Chua, N., Kay, S.: Circadian clock mutants in Arabidopsis identified by luciferase imaging. Science 267(5201), 1161–1163 (1995)

Mizuno, T., Nomoto, Y., Oka, H., Kitayama, M., Takeuchi, A., Tsubouchi, M., Yamashino, T.: Ambient Temperature Signal Feeds into the Circadian Clock Transcriptional Circuitry Through the EC Night-Time Repressor in Arabidopsis thaliana. Plant and Cell Physiology 55(5), 958–976 (2014)

Mockler, T.C., Michael, T.P., Priest, H.D., Shen, R., Sullivan, C.M., Givan, S.A., McEntee, C., Kay, S.A., Chory, J.: The DIURNAL project: DIURNAL and circadian expression profiling, model-based pattern matching, and promoter analysis. Cold Spring Harbor symposia on quantitative biology 72, 353–63 (2007)

Mucha, P.J., Richardson, T., Macon, K., Porter, M.A., Onnela, J.P.: Community structure in time-dependent, multiscale, and multiplex networks. Science 328(5980), 876–8 (2010)

Nagel, D.H., Kay, S.A.: Complexity in the wiring and regulation of plant circadian networks. Current Biology 22(16), R648–R657 (2012)

Nepusz, T., Yu, H., Paccanaro, A.: Detecting overlapping protein complexes in protein-protein interaction networks. Nature methods 9(5), 471–472 (2012)

Newman, M.E.: The structure and function of complex networks. SIAM review 45(2), 167–256 (2003)

Ni, Z., Kim, E.D., Ha, M., Lackey, E., Liu, J., Zhang, Y., Sun, Q., Chen, Z.J.: Altered circadian rhythms regulate growth vigour in hybrids and allopolyploids. Nature 457(7227), 327–31 (2009)

Noordally, Z.B., Ishii, K., Atkins, K.A., Wetherill, S.J., Kusakina, J., Walton, E.J., Kato, M., Azuma, M., Tanaka, K., Hanaoka, M., Dodd, A.N.: Circadian Control of Chloroplast Transcription by a Nuclear-Encoded Timing Signal. Science 339(6125), 1316–1319 (2013)

Nozue, K., Covington, M.F., Duek, P.D., Lorrain, S., Fankhauser, C., Harmer, S.L., Maloof, J.N.: Rhythmic growth explained by coincidence between internal and external cues. Nature 448(7151), 358–61 (2007)

Nusinow: A Diurnally Regulated ELF4-ELF3-LUX Complex is Critical for Circadian Control of Hypocotyl Growth. Nature 6 (2011)

Papana, A., Ishwaran, H.: CART variance stabilization and regularization for high-throughput genomic data. Bioinformatics 22(18), 2254–61 (2006)

Penfold, C.A., Buchanan-Wollaston, V., Denby, K.J., Wild, D.L.: Nonparametric Bayesian inference for perturbed and orthologous gene regulatory networks. Bioinformatics 28(12), i233–i241 (2012)

Ponnapalli, S.P., Saunders, M.A., Van Loan, C.F., Alter, O.: A higher-order generalized singular value decomposition for comparison of global mRNA expression from multiple organisms. PloS one 6(12), e28072 (2011)

Poyatos, J.: On the Search for Design Principles in Biological Systems 751, 183–193 (2012)

Priest, H.D., Filichkin, S.a., Mockler, T.C.: Cis-regulatory elements in plant cell signaling. Current opinion in plant biology 12(5), 643–9 (2009)

Radivojac, P., Clark, W.T., Oron, T.R., Schnoes, A.M., Wittkop, T., Sokolov, A., Graim, K., Funk, C., Verspoor, K., Ben-Hur, A., Pandey, G., Yunes, J.M., Talwalkar, A.S., Repo, S., Souza, M.L., Piovesan, D., Casadio, R., Wang, Z., Cheng, J., Fang, H., Gough, J., Koskinen, P., Törönen, P., Nokso-Koivisto, J., Holm, L., Cozzetto, D., Buchan, D.W.A., Bryson, K., Jones, D.T., Limaye, B., Inamdar, H., Datta, A., Manjari, S.K., Joshi, R., Chitale, M., Kihara, D., Lisewski, A.M., Erdin, S., Venner, E., Lichtarge, O., Rentzsch, R., Yang, H., Romero, A.E., Bhat, P., Paccanaro, A., Hamp, T., Kaß ner, R., Seemayer, S., Vicedo, E., Schaefer, C., Achten, D., Auer, F., Boehm, A., Braun, T., Hecht, M., Heron, M., Hönigschmid, P., Hopf, T.A., Kaufmann, S., Kiening, M., Krompass, D., Landerer, C., Mahlich, Y., Roos, M., Björne, J., Salakoski, T., Wong, A., Shatkay, H., Gatzmann, F., Sommer, I., Wass, M.N., Sternberg, M.J.E., Skunca, N., Supek, F., Bošnjak, M., Panov, P., Džeroski, S., Smuc, T., Kourmpetis, Y.A.I., van Dijk, A.D.J., Braak, C.J.F.T., Zhou, Y., Gong, Q., Dong, X., Tian, W., Falda, M., Fontana, P., Lavezzo, E., Di Camillo, B., Toppo, S., Lan, L., Djuric, N., Guo, Y., Vucetic, S., Bairoch, A., Linial, M., Babbitt, P.C., Brenner, S.E., Orengo, C., Rost, B., Mooney, S.D., Friedberg, I.: A large-scale evaluation of computational protein function prediction. Nature methods 10(3), 221–7 (2013)

Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., Barabási, A.L.: Hierarchical organization of modularity in metabolic networks. Science 297(5586), 1551–5 (2002)

Reikard, G.: Stimulating Economic Growth Through Technological Advance. Amstatnews 3 (2011)

Reshef, D.N., Reshef, Y.a., Finucane, H.K., Grossman, S.R., McVean, G., Turnbaugh, P.J., Lander, E.S., Mitzenmacher, M., Sabeti, P.C.: Detecting novel associations in large data sets. Science 334(6062), 1518–24 (2011)

Sanchez, S.E., Petrillo, E., Beckwith, E.J., Zhang, X., Rugnone, M.L., Hernando, C.E., Cuevas, J.C., Godoy Herz, M.A., Depetris-Chauvin, A., Simpson, C.G., Brown, J.W.S., Cerdán, P.D., Borevitz, J.O., Mas, P., Ceriani, M.F., Kornblihtt, A.R., Yanovsky, M.J.: A methyl transferase links the circadian clock to the regulation of alternative splicing. Nature 468(7320), 112–6 (2010)

Schmitt, W.A., Raab, R.M., Stephanopoulos, G.: Elucidation of gene interaction networks through time-lagged correlation analysis of transcriptional data. Genome Research 14(8), 1654–1663 (2004)

Searls, D.B.: Omic Empiricism. Science signaling 2(68), eg6 (2009)

Sharkhuu, A., Narasimhan, M.L., Merzaban, J.S., Bressan, R.A., Weller, S., Gehring, C.: A red and far-red light receptor mutation confers resistance to the herbicide glyphosate. The Plant Journal 78(6), 916–926 (2014)

Shen, D., Shen, H., Bhamidi, S., Muñoz Maldonado, Y., Kim, Y., Marron, J.S.: Functional Data Analysis of Tree Data Objects. Journal of Computational and Graphical Statistics 23(2), 418–438 (2014)

Slusarczyk, A.L., Lin, A., Weiss, R.: Foundations for the design and implementation of synthetic genetic circuits. Nature Reviews Genetics 13(6), 406–420 (2012)

Sprinzak, D., Elowitz, M.B.: Reconstruction of genetic circuits. Nature 438(7067), 443–8 (2005)

Stark, C., Breitkreutz, B.J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M.S., Nixon, J., Van Auken, K., Wang, X., Shi, X., et al.: The BioGRID interaction database: 2011 update. Nucleic acids research 39(suppl 1), D698–D704 (2011)

Stolovitzky, G., Monroe, D., Califano, A.: Dialogue on reverse-engineering assessment and methods: the DREAM of high-throughput pathway inference. Annals of the New York Academy of Sciences 1115, 1–22 (2007)

Straume, M.: DNA microarray time series analysis: automated statistical assessment of circadian rhythms in gene expression patterning. Methods in enzymology 383, 149–66 (2004)

Strumbelj, E., Kononenko, I.: An Efficient Explanation of Individual Classifications using Game Theory. Journal of Machine Learning Research 11, 1–18 (2010)

Swamidass, S.J., Azencott, C.A., Daily, K., Baldi, P.: A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval. Bioinformatics 26(10), 1348–56 (2010)

Tan, D.X., Manchester, L.C., Helton, P., Reiter, R.J.: Phytoremediative capacity of plants enriched with melatonin. Plant signaling & behavior 2(6), 514–6 (2007)

Terada, A., Okada-Hatakeyama, M., Tsuda, K., Sese, J.: Statistical significance of combinatorial regulations. Proceedings of the National Academy of Sciences 110(32), 12996–13001 (2013)

Tompa, M., Li, N., Bailey, T.L., Church, G.M., De Moor, B., Eskin, E., Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J., Makeev, V.J., Mironov, A.A., Noble, W.S., Pavesi, G., Pesole, G., Régnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C., Zhu, Z.: Assessing computational tools for the discovery of transcription factor binding sites. Nature biotechnology 23(1), 137–44 (2005)

Ukai-Tadenuma, M., Kasukawa, T., Ueda, H.R.: Proof-by-synthesis of the transcriptional logic of mammalian circadian clocks. Nature Cell Biology 10(10), 1154–1163 (2008)

Ukai-tadenuma, M., Yamada, R.G., Xu, H., Ripperger, A., Liu, A.C.: Delay in Feedback Repression by Cryptochrome 1 Is Required for Circadian Clock Function (2011)

van Ooijen, G., Millar, A.J.: Non-transcriptional oscillators in circadian timekeeping. Trends in biochemical sciences 37(11), 484–92 (2012)

Wang, P.I., Hwang, S., Kincaid, R.P., Sullivan, C.S., Lee, I., Marcotte, E.M.: RIDDLE: Reflective diffusion and local extension reveal functional associations for unannotated gene sets via proximity in a gene network. Genome biology 13(12), R125 (2012)

Wang, W., Barnaby, J.Y., Tada, Y., Li, H., Tör, M., Caldelari, D., Lee, D.u., Fu, X.D., Dong, X.: Timing of plant immune responses by a central circadian regulator. Nature 470(7332), 110–114 (2011)

Westmark, C.J.: A hypothesis regarding the molecular mechanism underlying dietary soy-induced effects on seizure propensity. Frontiers in Neurology 5(169) (2014)

Woller, A., Gonze, D., Erneux, T.: Strong feedback limit of the Goodwin circadian oscillator. Phys. Rev. E 87, 032722 (2013)

Xu, W., Yang, R., Li, M., Xing, Z., Yang, W., Chen, G., Guo, H., Gong, X., Du, Z., Zhang, Z., Hu, X., Wang, D., Qian, Q., Wang, T., Su, Z., Xue, Y.: Transcriptome phase distribution analysis reveals diurnal regulated biological processes and key pathways in rice flag leaves and seedling leaves. PloS one 6(3), e17613 (2011)

Yerushalmi, S., Yakir, E., Green, R.M.: Circadian clocks and adaptation in Arabidopsis. Molecular ecology (2011)

Yosef, N.: Dynamic regulatory network controlling Th17 cell differentiation. Nature 496(7446) (2013)

Zhang, B., Horvath, S., et al.: A general framework for weighted gene co-expression network analysis. Statistical applications in genetics and molecular biology 4(1), 1128 (2005)

Zoppoli, P., Morganella, S., Ceccarelli, M.: TimeDelay-ARACNE: Reverse engineering of gene networks from time-course data by an information theoretic approach. BMC Bioinformatics 11(1), 154 (2010)

Zou, M., Conzen, S.D.: A new dynamic Bayesian network (DBN) approach for identifying gene regulatory networks from time course microarray data. Bioinformatics 21(1), 71–79 (2005)