Engineering C₄ photosynthetic regulatory networks

Sarit Weissmann and Thomas P Brutnell

C₄ photosynthesis is a complex metabolic pathway responsible for carbon fixation in major feed, food and bioenergy crops. Although many enzymes driving this pathway have been identified, regulatory mechanisms underlying this system remain elusive. C₄ photosynthesis contributes to photosynthetic efficiency in major bioenergy crops such as sugarcane, Miscanthus, switchgrass, maize and sorghum, and international efforts are underway to engineer C₄ photosynthesis into C₃ crops. A fundamental understanding of the C₄ network is thus needed. New experimental and informatics methods can facilitate the accumulation and analysis of high-throughput data to define components of the C₄ system. The use of new model plants, closely related to C₄ crops, will also contribute to our understanding of the mechanisms that regulate this complex and important pathway.

Address
Boyce Thompson Institute for Plant Research, Cornell University, Tower Road, Ithaca, NY 14853, United States

Corresponding author: Brutnell, Thomas P (tpb8@cornell.edu)

Introduction

C₄ photosynthesis is an example of a complex metabolic pathway that has become an attractive target for pathway engineering in crop plants [1]. It drives the high rates of biomass accumulation of some promising bioenergy feedstocks including sugarcane, sorghum, Miscanthus, switchgrass and maize. Under hot arid conditions, C₄ crops have a higher yield potential, relative to C₃ plants, owing to a greater photosynthetic conversion efficiency [2] and increased water and nitrogen use efficiencies. Importantly, the C₄ pathway has evolved independently in at least 60 lineages of angiosperms, and at least 17 times in the grasses alone [3]. Thus, the evolutionary history of C₄ photosynthesis suggests that it is a plastic process, and that a biochemical foundation exists from which to engineer C₄ traits into C₃ crops.

A key element of the C₄ pathway is the coordinated metabolic activities of two distinct, specialized leaf cell types, mesophyll (ME) and bundle sheath (BS) (Figure 1). These two cell types function to shuttle CO₂ into the BS, effectively eliminating photosynthetic losses associated with photorespiration [4]. A comparative analysis of several carbon shuttle enzymes (Figure 1) suggests strong positive selection acting on duplicated gene family members resulting in alterations in the promoters and coding regions of these genes [5]. It appears that genomes of many C₄ plants have retained the ancestral C₃ functions of the C₄ enzymes through gene duplication events and subfunctionalization [6].

Spatial and temporal expression of carbon shuttle enzymes is probably mediated largely through transcriptional controls [10]. Fine-tuning of this process, however, requires additional levels of post-transcriptional, translational and post-translational control [12]. Although cis-regulatory regions have been defined for some genes encoding the C₄ carbon shuttle enzymes, no trans-acting regulators have been defined that control the high levels of cell-specific gene expression associated with C₄ photosynthesis.

C₄ photosynthesis regulatory sequences

Transcriptional regulatory networks are molecular systems in which environmental or developmental signals are integrated and transduced into differential gene expression. Transcriptional regulation is achieved by the combinatorial interplay of trans-acting protein complexes, and cis-regulatory sequence elements located in or near target genes [13]. Cis-regulatory sequences include promoters, enhancers, silencers, and insulators [14]. Promoters can be generally defined as consisting of a core promoter and a proximal promoter. The core promoter is located immediately 5’ to the transcription start site and serves as a binding site for the RNA polymerases II complex. The proximal promoter region is generally located within 1 kb of the core promoter, and includes sequence-specific transcription factor binding sites [13]. General transcription factors binding
the proximal promoter stabilize the binding of the RNA polymerases II complex to the core promoter and thus promote transcription [15]. Enhancers promote transcription by recruiting histone-modifying enzymes that open the chromatin structure, and thus increase the accessibility of the promoter [16]. Silencers can either actively interfere with the assembly of the pre-initiation complex, via the transcription factors that bind them, or passively prevent the binding of TFs to their respective cis-regulatory motifs [14]. Insulators are sequences that have the ability to protect genes from inappropriate signals emanating from their surrounding genomic environment [17]. They can also act as barriers, by preventing the advance of nearby condensed chromatin that might silence the expression of the gene [18].

To date, only a few transcriptional regulatory sequences that confer differential BS/M expression to C₄ enzymes have been described [12]. These include the phosphoenolpyruvate carboxylase (PEPC) proximal promoter and cis-element [8,19]. 5’- and 3’-noncoding sequences surrounding the NADP-ME gene [20–22], coding regions of NAD-ME [23*], the two promoter regions of pyruvate orthophosphate dikinase (PPDK) [24–27], and 5’ and 3’ sequences of RuBisCO [28–31]. The only sequence, however, that has been defined at nucleotide resolution, is the MEM1 element that confers mesophyll specificity of the PEPC in C₄ Flaveria [8]. As additional genome and transcriptome sequencing efforts proceed, there is a need for large-scale analysis of regulatory sequences that are involved in
mediating the differential and high level of C₄ gene expression.

C₄ photosynthesis trans-acting factors

The binding of transcription factors to specific DNA target sequences forms the foundation of transcriptional networks. Transcription factors recruit the RNA polymerase complex to the transcription start site, by binding to either sequences in the proximal promoter or to distant cis-regulatory sequences. Transcription factors may also interact with chromatin remodelers or modifiers that facilitate access or increase protein–protein affinities via histone modifications [32,33].

Surprisingly, little is known about the transcriptional network that regulates C₄ gene expression. Dof1 (DNA Binding with One Finger 1) and Dof2 have been shown to bind the maize PEPC1 and PPDK1 promoters and combinatorially regulate their expression [34,35]. It is assumed that Dof1 promotes the expression of PEPC, and that Dof2 represses the expression of Dof1 in non-photosynthetic tissues. Other transcription factors such as FtHB1 [36], MNFs (Maize Nuclear Factors) and PEP-I [37], were also shown to bind the promoter of PEPC. It is not clear, however, what precise role they play in the regulation of PEPC expression. For example, deletion of the FtHB1 binding site within the PEPC 5'UTR, does not eliminate M cell specific expression [38]. The Golden2 (G2) gene, encoding a GARF transcription factor [39,40], has been suggested as a major regulator of BS photosynthetic development. bsd1 mutants show cell-type specific aberrant chloroplast development in maturing BS cells, resulting in chlorotic regions in the leaf. In addition transcripts encoding enzymes associated with C₄ photosynthesis are lower in BS cells while showing normal transcript levels in M cell [41]. Interestingly, ectopic expression of the rice G2 homolog promotes plastid development in bundle sheath cells of rice [42]. However, the targets of G2 in maize and rice remain elusive.

Dissecting C₄ photosynthesis transcriptional regulatory networks

Dissecting transcriptional regulatory networks requires a thorough understanding of the complex spatial (e.g. tissue-specific) and temporal (e.g. developmental series) interactions between multiple transcription factors and gene targets. This global view of network function has been facilitated in recent years through the availability of sequenced genomes, new technologies (e.g. microarrays, RNA-seq), and informatics tools that allow the capture and interrogation of large datasets [43,44]. Combining these algorithms with genome-wide high-throughput experimental data, it is now possible to develop probabilistic Bayesian networks and Gaussian graphical models for regulatory and cellular networks [43,45]. To complement these methods, evolutionary data can be incorporated into the analysis. Selective pressures act on both coding and non-coding DNA sequences. In fact, in some cases, non-coding sequences can be subjected to a higher selection pressure than coding regions [46,47]. Phylogenetic footprinting assumes that conserved non-coding sequences have a high chance of being functional regulatory elements [48,49]. This technique has been widely used for large-scale identification of putative regulatory sequences in several species [50–54]. Sequence conservation analysis can be applied using algorithms, such as VISTA [55] and CoGe [56].

Additional biochemical techniques such as whole genome identification of regulatory sequences can be achieved by Chromosomal Conformation Capture [57,58]. This experimental method can identify in vivo physical interactions between genomic segments. It has been used to identify and confirm cis-regulatory elements [59,60], and to uncover potential genome-wide interactions between human conserved non-coding sequences [61]. Another useful way to reveal active chromatin is through DNase-seq (a high-throughput DNase I hypersensitivity assay). This method identifies exposed, and thus active DNA segments, and was used to map genome wide putative cis-elements in rice [62]. The knowledge of the location of the regulatory sequences should be combined with data on which TFs bind them, and which other genes are regulated by these TFs. An efficient way to test if the putative cis-regulatory element is bound by proteins is the electrophoretic mobility shift assay (EMSA). This method can identify protein–DNA binding via their migration on a gel, and can also be used to isolate the TFs that bind the regulatory sequence [63]. Another way to identify the specific TFs that bind a regulatory sequence is the yeast one hybrid (Y1H) assay. This method identifies protein–DNA interaction in vivo in yeast cells. This assay allows large-scale library screens [64], and was used to identify TFs involved in circadian clock regulation in Arabidopsis [65]. Ideally, confirmation of binding sites should be validated in vivo through the construction of reporter gene reporters. Once TFs are identified, their downstream targets can be defined using methods like ChIP-seq that allows the identification of all the sequences to which a TF binds [66] and by examining the consequences of ectopic expression and through mutant analysis. Cytosine DNA methylation also plays a crucial role in the stable expression of genes in plants [67]. High resolution mapping of methylated cytosines and small RNAs can be achieved by using high-throughput sequencing techniques [68]. Through an integration of these various datasets in a systems approach, it should be possible to establish a deep understanding of the regulatory mechanisms that underlie this complex process (Figure 2).
Gene expression is controlled at several levels of regulation. Here, we have highlighted transcriptional (green) and post-transcriptional (purple) levels. Stacking high-throughput datasets (RNA-seq, ChiP-seq, Bisulfite-Seq), obtained by experimental and computational methods (red), allows the construction of complex regulatory networks.

The need for new grass model systems

One of the great limitations in the functional dissection of transcriptional networks in C₄ plants has been the lack of suitable genetic model systems. Ideally, a C₄ model system should have a sequenced genome, small stature, short life cycle, simple growth requirements, be readily transformable and be evolutionary close to grass crops. Major grass crop species are evolutionary distant from the widely used model species Arabidopsis thaliana [69] and from other C₄ models such as Flavaria sp and Cleome sp. [70]. Thus, in recent years several groups have looked to new model plants that are evolutionarily closer to grass crops [71*,72–75]. Two monocot species have been recently emerging as attractive genetic models for the grasses: Brachypodium distachyon, and Setaria sp (S. viridis and S. italica) [71*,76,77**]. B. distachyon is a temperate C₃ grass with a sequenced genome [78]. It will probably serve as an excellent model for temperate pooid grasses such as wheat, rye and barley. Both S. viridis and S. italica are C₄ grasses that are closely related to many of the major C₄ crops, such as maize, sugarcane, and sorghum, as well as to the biofuel feedstock switchgrass, and the invasive weed guinea grass (Panicum maximum). Setaria species have a small (510 Mb) and sequenced genome (http://www.phytozone.net/foxtailmillet.php); they are small in stature (10–15 cm) with a rapid life cycle of 6–9 weeks, and high seed production (13 000 seeds per plant) [75]. Foxtail millet is a minor food crop in China and India, and also includes some of the worst weed groups interfering with world agriculture (giant, green, yellow, knotroot, and bristly foxtail) [79]. Setaria species share many important traits with grass crops and weeds, such as self-compatibility, secondary seed dormancy and light requirements to induce germination. They also include many genotypes with various herbicide resistances, as well as drought and salt tolerance [79], and can be transformed using Agrobacterium mediated methods [71*].

Given the favorable attributes of Setaria sp. discussed above, it is likely that Setaria will emerge as the primary model system for understanding C₄ networks in the grasses. Transient, stable and protoplast transformation techniques will allow rapid dissection of promoter and cis-regulatory elements, and the manipulation of transcription factor expression. Brachypodium will be extremely useful as a proxy, for testing the transfer of the C₄ regulatory network into recalcitrant genetic systems such as wheat and rye. Together, Setaria and Brachypodium promise to greatly accelerate our understanding of the complex network that drives C₄ differentiation.

Conclusions

Engineering a C₄ photosynthetic system into a C₃ grass crop is a challenging task that will require systems level
understanding of the C₄ transciptional regulatory network. Although very little is known about the components of the C₄ pathway, recent advances in sequencing methodologies and computational tools promise to accelerate the construction of transcriptional networks. As resources for new model genetic systems progress, we will also see rapid progress in the functional dissection of these complex networks. Ultimately, these tools and technologies will drive the development of new crops to feed and fuel a rapidly growing global population.

Acknowledgement
We would like to thank the NSF for supporting this work through a grant to T.P.B. (IOS-0701736).

References and recommended reading
Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
- of outstanding interest


This paper provides one of the most comprehensive views of the plastid proteome in maize and identifies several novel innovations associated with C₄ photosynthesis.


This paper shows that cis-regulatory elements within coding regions of a gene can drive cell-specific gene expression in C₃ and C₄ plants.


47. Moses AM, Chiang DY, Kellis M, Lander ES, Eisen MB: Position specific variation in the rate of evolution in transcription factor binding sites. BMC Evolutionary Biology 2003, 3:.


This high throughput method allows the parallelization of yeast one-hybrid library screening to capture potentially thousands of interactions in an experiment.


74. Gressel J.: Arabidopsis is not a weed, and mostly not a good model for weed genomics; there is no good model for weed genomics. Weedy and Invasive Plant Genomics. Wiley-Blackwell; 2009.; pp. 25-32.


77. Doust AN., Kellogg EA., Devos KM., Bennetzen JL.: Foxtail Millet: a sequence-driven grass model system. Plant Physiology 2009, 149:137-141. This paper introduces the Setaria sp. as model systems for grass research and describes many of the attractive aspects of this model system.
