CHARACTERIZING THE PHYTOCHROME GENE FAMILY IN ORYZA SATIVA

by

AYNSLEY PEEL EASTMAN

(Under the Direction of Marie-Michèle Cordonnier-Pratt)

ABSTRACT

Phytochromes are encoded by a light-responsive gene family inherent to all plants and the cyanobacterial progenitor of their plasmids. The genome of the plant *Oryza sativa* (rice) contains a three-membered phytochrome gene family. The region of the genome containing each phytochrome genes was sequenced and subsequently annotated using a method developed here to identify the genes upstream and downstream of each phytochrome gene. A comparative analysis of the phytochrome-containing sequences revealed that, due to the age of the gene duplication events that gave rise to the multi-membered phytochrome gene family, no mechanisms of gene family expansion can be currently detected in the genome. Analysis of the putative promoters of the phytochrome genes revealed that light-regulatory elements were among the most abundantly identified type of regulatory motif within sequences conserved between phytochrome rice/sorghum orthologs.

INDEX WORDS: phytochrome, rice, *Oryza sativa*, DNA sequencing, genomics, annotation, bioinformatics, conserved non-coding sequences, promoters, comparative genomics, photomorphogenesis

CHARACTERIZING THE PHYTOCHROME GENE FAMILY IN ORYZA SATIVA

by

AYNSLEY PEEL EASTMAN

Bachelor of Science, College of Charleston, 1999

Bachelor of Arts, College of Charleston, 1999

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2005

CHARACTERIZING THE PHYTOCHROME GENE FAMILY IN ORYZA SATIVA

by

AYNSLEY PEEL EASTMAN

Major Professor:	Marie-Michèle Cordonnier
Committee:	Lee H. Pratt Claiborne V. Glover

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2005

DEDICATION

This thesis is dedicated, in its entirety, to Andrew Hamilton Eastman, without whose love and support the first page would ever have been written.

ACKNOWLEDGEMENTS

I would like to first thank and acknowledge the two people most responsible for my graduate career, Drs. Lee Pratt and Marie-Michèle Cordonnier. Your guidance and help have been invaluable. Marie-Michele, thank you for teaching me everything I know about computers, databases, and scripting; your dedication is inspiring. Lee, thank you for teaching me about phytochrome, how to write effectively, and the correct definition of a noun-modifier.

I would also like to thank Claiborne Glover, John McDonald, and David Puett for their guidance and support throughout my years at the University of Georgia. To Coach Chuck McCord, Mr. Hugh Hill, Dr. Kristen Krantzman, and Dr. Rick Heldrich, thank you for your years of support and academic encouragement.

To my family; R.L., Kemper, Don, Chris, Zack, and Bryan, thank you for your patience, encouragement and love. And finally, I would like to thank Mom and Dad for always telling me I could do and be whatever I wanted-even when you didn't understand what I was talking about.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

GLOBAL INTRODUCTION

Introduction

Phytochromes are encoded by a light-responsive gene family inherent to all plants and the cyanobacterial progenitor of their plasmids. The genome of the plant *Oryza sativa* (rice) contains a three-membered phytochrome gene family. The region of the genome containing each phytochrome genes was sequenced and subsequently annotated using a method developed here to identify the genes upstream and downstream of each phytochrome gene. A comparative analysis of the phytochrome-containing sequences revealed that, due to the age of the gene duplication events that gave rise to the multi-membered phytochrome gene family, no mechanisms of gene family expansion can be currently detected in the genome. Analysis of the putative promoters of the phytochrome genes revealed that light-regulatory elements were among the most abundantly identified type of regulatory motif within sequences conserved between phytochrome rice/sorghum orthologs.

Purpose of the Study

The project presented here can be organized into three main goals:

(I) Develop a system that can handle the appropriate level of detail necessary for an in-depth and accurate annotation of *PHY*[1]-containing sequences in order to be able to readily extract biologically useful information.

(II) Sequence the *PHY* and their flanking DNA in rice (*Oryza sativa*), a model monocot, and extensively annotate these sequences to determine what genomic features (genes, non-coding sequences, repeat sequences, etc), if any, are conserved between these three

[1] *PHY* = phytochrome gene, PHY = phytochrome protein. A specific *PHY* or PHY is designated by the letter(s) of that *PHY*/PHY or *PHY*/PHY subfamily.

regions of the rice genome and *PHY*-containing regions of any publicly available plant genome.

(III) Analyze, in detail, the putative upstream promoter regions of the three *PHY* and the promoters of sorghum *PHY* to identify conserved non-coding sequences between the putative promoters. Also, identify regulatory motifs within the conserved regions detected between the sorghum/rice *PHY* orthologs and light-regulatory motifs in the rice *PHY* promoters.

CHAPTER 2

LITERATURE REVIEW: THE PHYTOCHROME GENE FAMILY, GENE DUPLICATION, ANNOTATION, AND COMPARATIVE GENOMICS IN PLANTS

Phytochrome History

Plants not only employ solar radiation as an energy source to drive photosynthesis, but they also use differences in light quantity, quality, and directionality to determine the appropriate timing for various life cycles and developmental stages. For example, photoperiod, which is the ratio between day and night, determines when a plant flowers in response to seasonal changes in daily light duration. W. W. Garner and H. A. Allard, while working at the U.S. Department of Agriculture at the Arlington Experiment Farm, determined that flowering plants fall into at least two categories (Garner and Allard, 1933). Long-day plants (LDPs) require daylengths longer than some critical length to flower, and short-day plants (SDPs) are those that fruit or flower through a wide range of daylengths up to a critical length (Garner and Allard, 1933). It was later determined that long-day plants could be induced to flower by interrupting a long night with a brief pulse of red light and, contrastingly, the same exposure inhibited flowering in short-day plants (Hendricks, 1964). These effects were reversible and could subsequently be cancelled by a short burst of far-red light (Borthwick *et al*., 1952). It was also found that other photomorphogenic responses were susceptible to red and far-red light and were similarly photoreversible (Flint and McAlister, 1935; Meischke, 1936). These observations, along with others, led Warren Butler to try to identify a photosensory pigment that underwent photoreversible absorbance changes. He was successful in his attempts to isolate such a photoreceptor and termed it phytochrome (Butler *et al.*, 1959), meaning “plant color”.

Phytochromes in all higher plants appear to have the same basic structure. Phytochrome is a homodimer of two 124-kDa subunits (Pratt, 1982; Lagarias, 1985). The N-terminal region, approximately 600 amino acids long, covalently binds a linear tetrapyrrole chromophore to a conserved cysteine residue (Furuya and Song, 1994). This domain, responsible for initiating the

photoconversion of phytochrome, is highly conserved both within the subfamilies of phytochromes within a single organism, as well as among phytochromes of evolutionarily divergent organisms. The C-terminus among phytochromes is more variable and is required for signal transduction and homodimerization. Moreover, the C-termini of phytochromes exhibit sequence homology to the histidine kinases of bacterial two-component regulatory systems (Schnieder-Poetsch, 1998, reviewed by Elich and Chory, 1997; Quail, 1997).

Phytochrome exists in two interconvertible forms. It is synthesized in darkness as the inactive red-light absorbing form (P_r). Upon absorption of red light, it is converted to the biologically active far-red-light absorbing form (P_{fr}) (Furuya and Song, 1994). This photoconversion of the chromoprotein occurs via a series of intermediates resulting in the biologically active conformation of phytochrome (Kneip *et al*., 1999). P_r and P_{fr} have overlapping spectra except in the far-red region. Consequently, phytochrome exists as a ratio of P_r/P_{fr}. It is this ratio of the inactive/active forms of phytochrome that allow it to sense the quality of ambient light and respond appropriately. For example, in the etiolation response in Arabidopsis, when sunlight converts P_r into P_{fr}, phytochrome moves from the cytoplasm into the nucleus where it binds to phytochrome-interacting factor 3 (PIF3) (Ni *et al*., 1999). The phytochrome/PIF3 complex binds to the promoter regions of genes encoding certain transcription factors that, in turn, induce transcription of light-responsive genes. The complex dissociates and returns to the cytoplasm when exposed to far-red light, which converts phytochrome back into the P_r form (Ni *et al*., 1999).

Phytochrome Function

Members of the PHY[1] family have different photosensory and/or physiological functions and phytochrome photoconversion is responsible for the regulation of many developmental and photomorphogenic responses in plants (Quail, 1998; Whitelam, Patel, and Devlin, 1998; Deng and Quail, 1999). In Arabidopsis, the specific function of each member of the phytochrome family (A-E) has been analyzed using mutants deficient in each respective gene (Dehesh *et al*., 1993; Nagatani, Reed, and Chory, 1993; Parks and Quail, 1993; Whitelam *et al*., 1993; Reed *et al*., 1993; Aukerman *et al*., 1997; Davis *et al*., 1999; Devlin *et al*., 1999; Muramoto *et al*., 1999; Franklin *et al*., 2003). The action spectra of three categories of light-dependent reactions were determined for each mutant to find the modes of photoperception for the different phytochromes (Parks and Quail 1993; Shinomura *et al*., 1996, Franklin *et al*., 2003). The three categories of photoreactions analyzed were: very low fluence responses (VLFR), low fluence responses (LFR), and high-irradiation responses (HIR) (Smith, Whitelam, and McCormac, 1991; Kuno and Furuya, 2000).

On the whole, these studies determined that PHY have separate, yet sometimes overlapping functions in Arabidopsis. For example, PHYA regulates both the VLFR and the far-red light-mediated HIR. VLF light from 380-780 nm was shown to trigger the non-photoreversible PHYA-dependent seed germination response (Shinomura *et al*., 1996). In contrast, PHYB regulates LFR and reversibly switches responses on or off upon exposure to LF red and far-red light, respectively (Shinomura *et al*., 1996). Furthermore, PHYA mediates hypocotyl inhibition (Nagatani *et al*., 1993; Parks and Quail, 1993: Whitelam *et al*., 1993), blue light sensing (Neff and Chory, 1998), and floral promotion (Johnson *et al*., 1994) in far-red light

[1] *PHY* = phytochrome gene, PHY = phytochrome protein. A specific *PHY* or PHY is designated by the letter(s) of that *PHY*/PHY or *PHY*/PHY subfamily.

environments. Because PHYA rapidly degrades in the presence of high intensity red light it is often considered as an "antenna" that detects the small amount of light that penetrates through the soil, thereby allowing seedlings to respond and emerge from the ground (Yanovsky *et al*., 1995). Phytochromes B-E are light stable in Arabidopsis and are generally involved in responses to higher levels of red light (Hirschfeld *et al*., 1998; Whitelam *et al*., 1998). PHYB mediates both inhibition of hypocotyl elongation in red light (Koornneef *et al*., 1980; Somers *et al*., 1991; Reed *et al*., 1994) and shade avoidance responses (Somers *et al*., 1991; Smith and Whitelam 1997). Although the inhibition of hypocotyl elongation in red light is a well-documented PHYB-mediated response (Koornneef *et al*., 1980, Somers *et al*., 1991), PHYC, in the absence of other phytochromes, is a red light sensor that also inhibits hypocotyl elongation under high fluence rates of red light in a FR-reversible manner (Franklin *et al*., 2003). Contrastingly, it has also been indicated by mRNA expression studies in the dicot *Stellaria longipes* (common name: Meadow starwort) that *PHYC* may promote stem elongation and flowering under long day photoperiods (Li and Chinnappa, 2004). PHYD, which is phylogenetically closely related to PHYB, also mediates shade avoidance (Aukerman *et al*., 1997; Devlin *et al*., 1999) and interacts with cry1, a known cryptochrome (Hennig *et al*., 1999). Cryptochromes are plant photoreceptors known to play a role in photomorphogenic responses in response to blue light (Ahmad *et al*., 1998; Lin *et al*., 1998). This interaction suggests that PHYD may be involved in the modulation of blue-light photoperception (Hennig *et al*., 1999). PHYE is involved in the maintenance of rosette habit, shade avoidance, and light-mediated germination (Devlin *et al*., 1999).

PHY is also known to be directly involved in maintaining a plant's circadian behavior. The gene products involved in photosynthesis build up shortly before daybreak each day in anticipation of dawn, while leaves and flowers often close before dusk to protect themselves

from lower nighttime temperatures (Darwin, 1895; Enright, 1982; Millar and Kay 1991). Unlike some of the photomorphogenic responses described above, these actions are not regulated solely by responses to external light levels, as a plant kept in constant light will still exhibit the cyclical production of the photosynthetic machinery (Millar and Kay, 1991). The circadian clock, which maintains an approximately 24-hour rhythm, is what controls these processes, yet the circadian clock does not function in isolation of daily light cycles (Devlin, 2002). The clock must be reset daily to synchronize the organism with ever-changing day/night cycles and seasonal progressions (Devlin, 2002). In Arabidopsis, phytochrome, combined with the blue light-absorbing cryptochromes, has been shown to relay day/night signals to reset the clock each day (Somers *et al*., 1998). As expected, *PHYA* mutants show deficiencies in perceiving low fluence rates of red light, which is consistent with the proposed involvement of PHYA as a low fluence photoreceptor in seedling establishment, while *PHYB* mutant seedlings display a deficiency in the perception of high fluence red light (Somers *et al*., 1998). These responses demonstrate the capacity of plants to recruit photoreceptors in response to the large range of red light fluence rates as PHYA responds to red light fluence rates below those that are in the active range of the PHYB receptor. Phytochromes D and E also provide input to the circadian clock. PHYD and E display some functional redundancy with PHYB as they are activated in the absence of PHYB (Whitelam *et al*., 1998). This redundancy is understandable given that, in Arabidopsis, PHYB and PHYD are thought to have arisen as a result of a recent gene duplication event subsequent to the divergence of the Cruciferae and Solanaceae from a common ancestor (Mathews and Sharrock, 1997).

Some PHY are also known to interact with other proteins in signaling pathways. Several loci have been identified that are specific to either PHYA or PHYB signaling pathways via

genetic screens (Whitelam *et al*., 1993; Ahmad and Cashmore, 1996; Wagner *et al*., 1997; Hoecker *et al*., 1998; Soh *et al*., 1998; Hudson *et al*., 1999) or yeast two-hybrid experiments (Ni *et al*., 1998; Choi *et al*., 1999; Frankhauser *et al*., 1999). One of these loci, PIF3, produces a nuclear-localized, basic helix-loop-helix (bHLH) protein (Ni *et al*., 1998) that belongs to the bHLH superfamily of transcription factors (Atchley and Fitch, 1997; Littlewood and Evan, 1998). In the nucleus, a PHY/PIF3 complex binds to the promoter regions of genes either directly involved in photomorphogenic responses, or those involved in the light-regulation of these responses, suggesting the complex serves a direct role in the signaling pathway from photoreceptors PHYA or PHYB to a target gene (Kircher *et al.,* 1999; Yamaguchi *et al*., 1999). This hypothesis is supported by the fact that PHYA and PHYB are known to translocate from the cytoplasm to the nucleus after the active conformation of phytochrome is formed (Kircher *et al.,* 1999; Yamaguchi *et al*., 1999).

The functions of PHY and the involvement of PHY in signaling pathways described here have been determined through multiple experimental approaches such as mutagenesis and yeast two-hybrid experiments. Yet there is still much to discover and understand regarding the discrete and/or overlapping functions of PHY in various organisms, as well as PHY's involvement in the complex signaling mechanisms that allow a plant to respond to its surrounding environment. The advent of multiple plant genome projects and the production of DNA sequence information from multiple non-model plant species coupled with emerging bioinformatic techniques will facilitate the study of PHY interactions with other proteins and other gene promoter sequences. Understanding the structure and purpose of those mechanisms regulating PHY, and how those mechanisms interact with other regulatory controls, will increase our knowledge of PHY function and participation in plant development.

Phytochrome Evolution

A predominant question in phytochrome research is how and why did this family expand? The introduction of angiosperms to the terrestrial landscape fundamentally changed the distribution of the dominant plant lineages present in the ecosystem at the time of angiosperm inception between 140-190 Mya[2] (Sanderson and Doyle, 2001). It has been hypothesized that the first angiosperms were probably understory plants that occupied shady habitats in tropical rain forests (Arber and Parkin, 1907; Bessey, 1915; Bews, 1927; Axelrod, 1952; Takhatajan, 1969; Thorne, 1974; Cronquist 1988). In regards to their light environment, early angiosperms could only have benefited from any evolutionary modification that made them able to thrive and establish themselves in a dimly lit environment (Mathews, Burleigh, and Donoghue, 2003). Indeed, it has been recently hypothesized that having a photoreceptor like PHYA, which exhibits a high degree of sensitivity to light detection, might have imparted an adaptive advantage to early angiosperms (Mathews, Burleigh, and Donoghue, 2003).

Multiple lines of evidence support the existence of a single cyanobacterial phytochrome progenitor to all existing phytochrome genes. First, as the earliest evolved photosynthetic organisms (Tandeau de Marsac and Houmard, 1993), cyanobacteria are believed to be the ancestors of all plant chloroplasts via an endiosymbiotic event (Gray and Doolittle, 1982). Second, the genome sequencing project of *Synechocystis* PCC 6803 (Kaneko *et al*., 1996), a unicellular non-nitrogen-fixing cyanobacterium, revealed a gene whose translated amino acid sequence was similar to eukaryotic phytochromes along its entire length. This gene, subsequently named Cph1 for cyanobacterial phytochrome 1 (Yeh *et al*., 1997), showed sequence similarity to bacterial sensory kinases in its C-terminal region (Lamparter *et al*., 1997). Lastly, the protein produced by Cph1, Cph1p, is a light-regulated two-component histidine-

kinase that shows red/far-red light induced optical transitions between the P_{fr} and P_r states typical of angiosperm phytochromes. It had previously been established by Yeh and Lagarias (1998) that existing eukaryotic phytochromes are light-regulated serine/threonine kinases with histidine kinase ancestry (Yeh *et al*., 1998).

The exact history of an organism's phytochrome evolution is specific to each species and, therefore, cannot be easily described on a generic level for the plant kingdom as a whole. The most common generic hypothesis of *PHY* family expansion from its cyanobacterial progenitor submits that, prior to the divergence of conifers from angiosperms, but after the divergence of seed plants (Donoghue and Mathews, 1999), the progenitor *PHY* duplicated to yield the *PHYA/C/F* and *PHYB/D/E* subfamilies (Figures 2.1 and 2.2) (Mathews and Sharrock, 1996). The first gymnosperms (ginkos, conifers and cycads) did not appear in the fossil record until the early Triassic period ~ 250 Mya (Donoghue, 1994), and the duplication of the progenitor *PHY* to yield the *PHYA/C/F* and *PHYB/D/E* progenitors predates this event (Donoghue and Mathews, 1999). The *PHYA/C/F* subfamily then underwent a second duplication event before the monocot/dicot divergence in angiosperms to yield the *PHYA* and *PHYC/F* subfamilies (Donoghue and Mathews, 1999) ~200 Mya (Wolfe, *et al.*, 1989). This second duplication yields the three-membered *PHYA*, *PHYB*, and *PHYC* family characteristic of current-day monocots (genes denoted by * in figure 2.1). This *PHYA/C/F* to *PHYA* and *PHYC/F* divergence-hypothesis is supported by the fact that, while dicot *PHY* appear to be separated into four or five subfamilies (*PHYA*, *B*/*D*, *C*, *E*, and *F*) (Clack, *et al.*, 1994; Matthews and Sharrock, 1996), monocots have representatives from only *PHYA*, *PHYB*, and *PHYC* (Mathews and Sharrock, 1996). Increasing the complexity of the gene family and depending on the specific evolution of PHY within a particular plant species, *PHYB/D/E* can duplicate to yield the *PHYB/D* and *PHYE*

[2] Mya=million years ago

subfamilies and *PHYC/F* and *PHYB/D* can duplicate to yield *PHYC, PHYF, PHYB,* and *PHYD*, respectively.

The characterization of the phytochrome gene families in *Arabidopsis thaliana* (Sharrock and Quail, 1989; Clack *et al*., 1994), *Solanum lycopersicum* (tomato) (Pratt *et al*., 1997), and *Ceratodon purpureus* (moss) (Pasentsis *et al*., 1998), as well as other angiosperms (Howe *et al*., 1989) and non-angiosperms (Quail, 1991; Kolukisaoglu *et al*., 1993; Mathews, Lavin, and Sharrock, 1995; Mathews and Sharrock, 1996), has demonstrated that the phytochrome gene family has become substantially larger and more complex during its evolution from its cyanbacterial progenitor (Kolukisaoglu *et al*., 1993; Mathews and Sharrock, 1996) and that the exact nature of this evolution differs from species to species. In the dicots Arabidopsis and tomato, the *PHY* family has been extensively modeled and shown to have five phytochrome loci each (*PHYA*-*PHYE/A*-*F*) (Mathews and Sharrock, 1996; Sharrock and Quail, 1989; Clack *et al*., 1994, Hauser *et al*., 1995; Pratt *et al*., 1997). Interestingly, in tomato the *PHYF* sequence displays less identity to Arabidopsis *PHY* than to other members of the tomato *PHY* family. Consequently, it was put into a separate and new subfamily (Hauser *et al*., 1995; Pratt *et al*., 1997). Other important differences within the *PHY* gene families of angiosperms are known. Two *PHYA* genes have been found in carnation (*Dianthus caryophyllus*) and some legumes (Fabaceae) (Mathews, Lavin, and Sharrock, 1995). Apparently, these duplications occurred independently in these species (Mathews, Lavin, and Sharrock, 1995). Witchgrass (*Panicum capillare*), a monocot, is also known to have two *PHYA* genes (Matthews and Sharrock, 1996), and *PHYA* pseudogenes have been found in maize and pea (Christensen and Quail, 1989; Sato, 1990). Both carrot (*Daucus carota*) and potato (*Solanum tuberosum*) contain multiple members of the *PHYB/D* subfamily, much like tomato described previously (Mathews, Lavin, and

Sharrock 1995; Pratt *et al*., 1995, 1997). This individualized nature of the phytochrome gene family within specific plants leads to the necessity of characterizing this gene family within multiple species. One cannot simply rely on *Arabidopsis* as the sole model plant because of the differences and independent evolution and gene duplication of the *PHY* family among angiosperms.

Gene Duplication

The fate of duplicated genes has traditionally been described in one of two ways. First, degenerative mutations might silence the gene leaving an inactive pseudogene. Second, the duplicate gene might evolve to exhibit a novel and beneficial function and thus be preserved (Lynch and Conery, 2000; Lynch *et al*., 2001). In either hypothesis, the original gene retains its native function. Indeed, studies in humans and mice on the fate of duplicated genes revealed that 50% of duplicated genes lose their function after duplication, becoming pseudogenes, whereas the other 50% experience some kind of functional divergence (Nadeau *et al*., 1997). However, a third hypothesis has also emerged which suggests that, after duplication, both the native and duplicated genes might evolve to specialize their functions by using the duplication event to "fine-tune" some developmental or signaling responses (Lynch *et al.*, 2001). Regardless of whether the duplicated gene evolves into a gene whose protein product exhibits a novel function, or the two genes both develop more specialized responses, the occurrence of gene duplication inarguably imparts an opportunity for the increased fitness of an organism by increasing its genetic diversity.

Pseudogene studies in model eukaryotes such as yeast (Harrison, Echols, and Gerstein, 2001; Harrison *et al*., 2002a), Drosophila (Harrison *et al*., 2003), and human (Harrison *et al*.,

2002b; Zhang *et al*., 2002; Torrents *et al*., 2003) have shown that the increased formation of psuedogenes in a specific gene family can be correlated with genes in the family duplicating rapidly, causing the family to increase in number of members (Harrison and Gerstein, 2002). When 64 prokaryotic genomes were used to identify pseudogenes, the most notable and numerous pseudogenes detected were from large, divergent gene families, such as histidine kinase-like ATPases (Liu *et al*., 2004). This study suggests that large gene families that are evolving and duplicating rapidly may generate more pseudogenes than more slowly evolving gene families (Liu *et al*., 2004). The identification of pseudogenes as evidence of a gene duplication event in a genome, however, can be problematic as they are more difficult to detect in nucleotide sequence comparisons because of their high evolutionary rate.

Yet, *how* does gene duplication increase the fitness or functional diversity of an organism? In yeast (Winzeler *et al*., 1999), *Drosophila* (Cadigan, 1994), mouse (Saga, 1992), and other vertebrates (Gibson and Spring, 1998) the phenotype of an organism has been shown not to be particularly affected by deleting a duplicate gene. This outcome can be explained by two different methods of compensation taking place within the organism. First, alternative metabolic pathways and regulatory networks could step in to compensate for the gene loss (Gu, 2003). Second, if a gene for which a duplicate exists is deleted, the duplicate gene could compensate for the loss of the other copy. This situation would increase the fitness of an organism with a duplicate gene by adding a level of protection against gene loss over an organism with no such duplicate. It has also been demonstrated that the long-term retention of a duplicate gene is statistically significant and does not typically occur at random (Nowak *et al.*, 1997; Lynch *et al*., 2001).

The mechanisms by which genes duplicate to increase the potential genetic complexity of an organism are unclear. Not only do individual genes duplicate, but also blocks of genes duplicate as well. Even entire chromosomes or genomes are now thought to have contributed to the evolution of animals, fungi, and plants (Kent and Zahler, 2000; Bancroft, 2000; Dehal *et al*., 2001). How does one determine the mechanisms involved in the evolution of a gene family? Is it possible to distinguish whole genome duplication events from homologous recombination, chromosomal duplication, or specific gene duplication, loss, or transposition? The availability of genome sequences can greatly aid in the study of gene families within organisms. By identifying a number of genes within a family, the genetic diversification of each member can then be investigated to estimate how long ago the gene evolved from a common ancestor. Although many gene families have been studied without the benefit of the entire genome sequence, the availability of the genome sequence does allow for the potential identification of more distantly related members of a family of pseudogenes that might not be identifiable by more traditional methods. Furthermore, by comparing large genomic sequences to each other, one can look for evidence of possible mechanisms of divergence and gene family expansion. For example, if a large region of a chromosome duplicated, rather than a discrete gene, one would see synteny between the original and duplicated regions. However, gene loss, translocation, and evolutionary divergence over time might disrupt this conservation of gene order between the duplicated regions of the genome. Again, contributing to the difficulty in detecting conserved gene order and evidence of tandem duplication events within a genome is the observation that transferred and duplicated genes tend to have a higher chance than non-transferred genes of becoming pseudogenes (Liu *et al*., 2004), and thus become more difficult to detect based on sequence comparisons to known genes. This difficulty is compounded by the fact that as the duplication

events become more ancient and sequences more divergent, it is more difficult to find evidence of how they might have occurred.

How does gene duplication and the mechanisms of gene duplication relate to the expansion of the phytochrome gene family? Presumably two duplications occurred before the monocot/dicot split ~ 200 Mya (Wolfe, *et al.*, 1989) to yield a three-membered *PHY* family (Mathews and Sharrock, 1996). The three-membered family then further duplicated and expanded to yield the five-membered dicot *PHY* family and the individual members of species-specific *PHY* families, some of which are described above. The exact mechanism of these duplication events is unknown. The increased number of phytochrome genes in dicots could reflect the possibility that dicots possess additional responses to red light that the "extra" phytochromes mediate; alternatively, it could be that monocots can accomplish the same things with fewer photoreceptors. One method of determining whether these genes duplicated individually or with large regions of flanking DNA is to sequence the regions of the genome upstream and downstream of the individual *PHY* and compare these regions to each other to identify either regions of DNA similarity or conserved genes or pseudogenes. The age of the duplication events specific to phytochrome, however, might hinder any identification of what once might have been conserved genomic regions flanking the phytochromes. The pitfalls specific to large-scale genome comparisons of ancient and divergent species such as rice and Arabidopsis, which are separated by the ~200 million years of monocot/dicot evolution (Wolfe, *et al.*, 1989) will be more extensively reviewed in a later section. Yet, regardless of the investigation of gene duplication mechanisms, the first step in extensively characterizing the phytochrome gene family and its surrounding DNA in any plant species is to accurately and

completely identify genes within the appropriate genome sequences that surround *PHY* and complete a comprehensive annotation.

Gene Discovery in Genomic Sequences and Annotation

Traditionally, two main approaches to identifying individual genes within a genomic sequence exist (Borodovsky, Rudd and Koonin, 1994; for a comprehensive review of available methods see Mathé *et al*., 2002). Homology-based methods of gene discovery rely on the identification of sequence similarity between known gene sequences and the nucleotide sequence with unknown gene content. *Ab initio* methods of gene discovery are algorithms and programs that use an informatic and statistical approach to identify nucleotide sequences defined as containing a high probability of being likely to contain a gene. Both homology-based and *ab initio* approaches to gene identification have intrinsic advantages and disadvantages.

Homology-based gene identification methods such as BLAST (Altschul *et al*., 1990) and FASTA (Pearson and Lipman, 1988) compare nucleotide sequence of unknown gene content to data sets of known genes or proteins such as the EST database of the National Center for Biotechnology Information (NCBI's dbEST) (Boguski *et al*., 1993) and the Non-Redundant protein REFerence database of the Protein Information Resource (PIR-NREF) (Wu *et al*., 2003), respectively. The most obvious advantage to this method is that when a comparison yields a result of sequence similarity that is of high percentage in terms of sequence identity over the entire length of the known gene, and contains well-defined exon/intron boundaries, a convincing argument can be made that a particular gene does exist in that location on the DNA strand of unknown gene content. Moreover, if this comparison yields sequence similarity to a known annotated protein sequence, a putative function for that gene can be surmised. The two obvious

disadvantages to homology-based gene identification methods, not including problems that arise due to erroneous identity or function assignments in data sets of known genes or proteins, are that (1) only those genes that have previously been sequenced from either a closely related species or a highly conserved ortholog or paralog can be identified and (2) the exact identification of the gene structure can be difficult to identify because gene and protein sequences of similar identity might not be identical and exactly share all domains or be full gene or protein sequences. Also, untranslated regions are difficult to define using this approach, as they are not as highly conserved as gene coding sequences. These disadvantages can be compounded when employing EST databases because partial ESTs do not define overall gene structure and the assignation of an EST to an individual gene in a multi-gene family is not trivial (Mathé *et al*., 2002). The disadvantages to this approach are becoming ever smaller as more and more genome, cDNA, and protein sequences are being produced and stored in publicly available databases. For example, as of October 4, 2004, PIR-NREF (Wu *et al*., 2003) contained over 1.8 million entries (http://pir.georgetown.edu/pirwww/ search/pirnref.shtml) and dbEST (Boguski *et al*., 1993) contained almost 24 million entries (http://www.ncbi.nlm.nih.gov/dbEST/).

Ab initio or algorithmic methods of gene detection rely on a set of statistical values that define the probability of gene or coding region being present in a particular area of a genome. Nucleotide composition, or G+C content, hexamer frequency, and codon composition are all characteristics that might indicate the presence of a gene. For example, introns are more A/T-rich than exons, especially in plants (Goodall and Filipowicz, 1989). Many gene prediction algorithms, such as GeneMark (Borodovsky and McInich, 1993) and Genscan (Burge and Karlin, 1997), are based on Markov models. In order to produce a Markov model on which to base the probability of exon or gene presence, a training set of sequences is required from which

probabilities can be obtained. The algorithms are typically trained on data sets of known genes and ESTs, which can bias the algorithms in two ways. First, because known genes are typically used to train the algorithms, genes that occur at a more frequent rate will be discovered, or "known" more often and subsequently are more abundant in the training set. This leads to these types of genes being more frequently identified by the algorithm over as compared to less frequently identified genes. Second, EST sets are also used to train algorithms. Because the majority of ESTs present in a typical set are 5' ESTs, the 3' end of genes will usually be underrepresented in the training set. A comprehensive list of publicly available gene prediction programs sorted alphabetically by their source can be found at the Computational Gene Recognition web site (http://www.nslij-genetics.org /gene/programs.html).

Other, more global, problems for gene identification exist. Several tools and algorithms have been developed to address them (Mathé *et al*., 2002). Programs such as SIM4 (Florea *et al*., 1998) and Spidey (Wheelan, 2001) specifically address the improvement of genomic DNA alignments to EST or cDNA sequences, as these types of alignments are particularly problematic. In addition to homology-based and *ab initio* methods, many other programs facilitate the accurate prediction of gene structure. For example, several programs, one of which is DIALIGN (Morgenstern, 2000), exist that attempt to detect sequence similarity from genome/genome alignments from related organisms hypothesizing that coding DNA sequences are more conserved that non-coding. Programs such as RepeatMasker (http://ftp.genom.washington.edu/RM/RepeatMasker.html) eliminate or mask repetitive sequences that might interfere with gene prediction. All of these approaches must be combined to facilitate the accurate identification of genes within a genomic sequence.

It is commonly accepted that employing multiple prediction programs and homology searches to predict the presence and locations of genes within an unannotated sequence increases the accuracy of gene predictions (Rouzé *et al*., 1999). For these reasons an approach that combines traditional prediction algorithms and homology-based analysis is usually best for identifying the maximum number of genes in an unannotated sequence with the highest possible accuracy. This approach of combining multiple methods demands a robust system of data storage, comparison, and retrieval. There are many examples of this type of combinatorial approach. The need to consider multiple lines of evidence explains why so many database projects have arisen in response to genome annotation. These databases are usually, but not always, dedicated to a particular species or class of organism, and are typically very large and extremely complex.

For example, the Ensembl database project, currently the most widely used public whole-genome annotation system, is a comprehensive source of automatically annotated human, mouse, and other genome sequences that are available as flat files or at an interactive web site (Clamp *et al*., 2003). Ensembl data and code is freely available at hppt://www.ensemble.org, as the group developing Ensemble seeks to create portable systems with which large genomic sequences can be analyzed, stored, and visualized, and to distribute these systems worldwide. Ensembl now contains nine genomes: human, mouse, rat, fugu, zebrafish, *C. briggsae*, *C. elegans*, *D. melanogaster*, and *A. gambiae* (Birney *et al*., 2004). Ensembl also supports the Distributed Annotation System (Dowell *et al*., 2001), which allows a third-party user to combine and view local data in the context of the vast public data available through Ensembl. The software system is based on a MySQL relational database and uses Bioperl as a base bioinformatics library (http://www.bioperl.org/) with Perl Application Programming Interface (API) (Curwen *et*

al., 2004). The Ensembl gene-building system annotates supported genomes based on evidence compiled from known EST, cDNA, and protein sequences. The system is seeking to incorporate gene prediction based on comparative analyses to other related genomes to increase gene orthology resources as well as to include the annotation of nonprocessed pseudogenes and regulatory elements (Curwen *et al*., 2004). While this resource is extremely valuable to the community at large, it can not be realistically installed and efficiently adapted by an individual annotator seeking to analyze in-depth a limited genomic region, especially for a genome not supported by Ensembl.

Specifically in regards to grass and rice gene discovery and annotation, there are two main resources for the storage, annotation, organization, and display of the massive amounts of data that have been generated. Gramene (Ware *et al*., 2002) and RiceGAAS (Sakata *et al*., 2002) are considered to be two of the most comprehensive publicly available plant databases.

Gramene (Ware *et al*., 2002) gathers information and data regarding rice and other members of the grass family and stores and organizes it in a manner that can be interactively viewed via a web browser. For example, the following features are mapped onto an interactive graphical representation of a rice genomic sequence: EST and unigene sequences from various grasses, gene models as predicted by FGeneSH (http://www.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind), rice full and partial cDNA sequences, and the annotation submitted to GenBank by the producer of the sequence. Yet there is no assignment of a final annotation. Like Gramene, the RiceGAAS (Sakata *et al*., 2002) automatic annotation system, which was developed by the Rice Genome Project (RGP), does not currently incorporate manual curation of data. RiceGAAS is the predominant method of annotation used by the International Rice Genome Sequencing Project (IRGSP). It identifies features within genome

sequences through a series of homology-based and *ab initio* gene predictors, along with other programs, combined with an algorithm designed to score the probability of the existence of each exon. Results are displayed graphically in a web browser. This scoring algorithm is designed to mimic the practices of a manual annotator. While a database containing the annotated rice sequences submitted to GenBank has been made public, the RiceGAAS annotation system itself is not publicly available.

Ensembl, Gramene, and RiceGAAS are all important databases for the genome community at large. They do not necessarily, however, address the concerns of a medium- to low-throughput genome sequencer seeking to semi-manually annotate a specific portion of a genome of interest. A tool that is highly-customizable to the annotation of *PHY*-containing genomic regions must be developed for the specific investigation of genomic entities such as pseudogenes, *PHY* orthologs and paralogs, and evidence of evolutionary gene duplication mechanisms such as transposons and ancient *PHY*.

Comparative Genomics in Plants

Comparisons of entire genomes are, among other things, beneficial to the timely annotation of new genomes and the increased understanding of a genome's evolutionary history. The variety of tools available for these large-scale comparisons reflects the emergence of this technique. The understanding of inter- and intra-genomic similarities within a genome and between a genome and the genomes of closely related species is necessary for both annotation and the study of genomic evolutionary history. When approaching plant comparative genomics specifically, however, the large evolutionary distance separating model monocots and dicots, such as

Arabidopsis and rice, creates problems not necessarily addressed by mammalian or prokaryotic comparative approaches.

The first comparative mapping studies performed on the grasses revealed a high degree of conservation of the map position and colinearity of many markers and quantitive trait loci (QTL) between different grass genomes, despite the up to 40-fold difference in genome size characteristic of grass genomes (Paterson *et al*., 1995; Pereira and Lee, 1995; Devos and Gale, 1997; Gale and Devos, 1998; Keller and Feuillet, 2000; Wang *et al.*, 2001). QTL imparting important agronomic traits like dwarfing were also found to be colinear among grass species (Paterson *et al*., 1995; Pereira and Lee, 1995; Wang *et al.*, 2001). Contrastingly, macro-colinearity has not been revealed between wheat and rice. Wheat is a polyploid whose genome is 40 times larger than that of rice (Arumuganathan and Earle, 1991). Moreover, its genome consists of 25-30% duplicated genes and 80% repetitive sequence. Some regions in barley (Dunford *et al*., 1995) and wheat (Yan *et al*., 2003) have been previously shown to demonstrate conserved gene order or micro-colinearity. In contrast, when 5780 wheat ESTs were compared to 3280 ordered rice BAC/PAC clones, it was determined that macro-colinearity is not particularly conserved between rice and wheat as substantial chromosomal rearrangements were revealed between markers (La Rota and Sorrells, 2004). Nevertheless, the lack of macro-colinearity found in this sequence-based comparative mapping approach does not necessarily mean that micro-colinearity is not preserved between rice and wheat at some points within the genome. The differing degrees of macro-colinearity displayed by comparative mapping studies leads to the question of whether this conservation within the genome is or is not displayed at the molecular level.

The first study to analyze the question of whether colinearity between grasses can be seen at the molecular level was performed by restriction mapping and partially sequencing the genomic regions containing the sh2/a1 locus in maize, sorghum, and rice (Chen *et al*., 1997). This study demonstrated that gene order was conserved between these species in this area of the genome. A second study revealed that microcolinearity also exists between wheat and barley genomes surrounding the loci encoding orthologous receptor-like kinases (Feuillet and Keller, 1999). However, other studies reveal significant lapses in microcolinearity between otherwise conserved genomes. Sequence comparisons between maize, sorghum, and rice show significant gene rearrangements, deletions, and insertions surrounding the Adh1 locus that have occurred during the ~ 50 million years of evolution separating these organisms (Tikhonov *et al*., 1999, Tarchini *et al*., 2000). Moreover, a similar lack of microcolinearity has been described surrounding the stem rust resistance gene rpg1 in barley and rice (Kilian *et al*., 1997).

Previous publications have proposed different hypotheses regarding the amount of conservation one could expect when comparing genomes with such large divergence dates, as would be characteristic when comparing grasses, monocots and dicots, or other divergent plant species. One model estimated that 50% of the genes within a 3 cM region would be maintained by chance between monocot and dicot genomes (Paterson *et al*., 1996). A second study proposes that detected regions of conservation will be < 10 cM in length when Arabidopsis, a dicot, is compared to genomes that diverged from a common ancestor > 100 Mya (Vision *et al*., 2000). A third study concludes that any macro-synteny that once existed between the monocot and dicot species is long eroded and no longer detectable using a comparative mapping strategy (Devos *et al*., 1999). Furthermore, these authors suggest that conserved micro-synteny can exist where macro-synteny is absent, because hundreds of genes might lie between adjacent markers.

Therefore, the absence of macro-synteny might not be indicative of a lack of micro-synteny. In contrast, but following the same logic, macro-synteny might be observed between two species when micro-synteny is absent.

Two studies specifically have addressed this question of conservation across the monocot/eudicot divide using two similar, yet distinctly different approaches. The first study utilized BLASTN (Altschul *et al*., 1990). This study revealed that within two regions of the rice and Arabidopsis genomes, Arabidopsis chromosome IV and rice chromosome II there were five regions of sequence homology. When the partial sequence of rice chromosome II was analyzed, 56 putative genes that might encode a protein product were discovered (Mayer *et al*., 2001). The five regions of conservation identified between rice and Arabidopsis contained 39% (or 22) of the 56 putative coding genes initially identified in the annotation of the rice sequence, suggesting a relatively high conservation level across large portions of these regions of the genomes (Mayer *et al*., 2001). This BLASTN-based strategy has been criticized, however, because the algorithm might align a sequence with a conserved domain within a non-orthologous gene or within separate members of a multigene family, rather than actual gene orthologs (Devos *et al*., 1999).

In contrast, when a second portion of the rice genome was compared to the translated annotated coding sequences (CDSs) of Arabidopsis using BLASTX (Altschul *et al*., 1997) only 60 small-scale regions of synteny between the genomes were identified (Salse *et al*., 2002). Each region contained 4 to 22 annotated genes. The authors suggest that this limited level of microcolinearity between Arabidopsis and rice will probably result in the inability to identify agronomic traits in the cereals using the Arabidopsis genome and map-based cloning. The development of this method was also an attempt to avoid the criticisms leveled at the BLASTN approach (Salse *et al*., 2002). For example, the detection of repetitive domains within a protein

was avoided by limiting the overlap of each hit to fewer than 15 amino acids. Furthermore, tandem duplication events in the rice genome were described as neighboring rice genes that matched to the same Arabidopsis CDS and, to avoid problems that might arise due to differences in genome size and gene density between Arabidopsis and rice, distances between hits within a syntenic block were measured in gene numbers, not in base pairs. After analyzing the validity of syntenic regions by a permutation statistics test, it was determined that syntenic regions with four or more gene members are highly significant. These results suggest that the conservation level between Arabidopsis and rice is very low (~ 17% gene conservation in homeologous segments), which contrasts with previous studies (*e.g.*, 50%; see Paterson *et al*., 1996, and 39%; see Mayer *et al*., 2001). The fact that the earlier studies did not address the existence of multi-gene families and tandem duplication events might explain the higher percentages of detected sequence homology reported.

Taken together all of these studies reveal the difficulty of detecting conserved regions of genomes throughout the grass family, as well as across the monocot/eudicot divide. Moreover, detecting the mechanisms of gene family expansion, evolution, and divergence becomes increasingly more difficult the further a species gets from the origin of the original duplication event that yielded the multi-membered gene family. Regardless of the ability to ultimately and successfully identify evidence of genome or gene family evolution, the initial task in any large-scale sequence comparison study is to generate the necessary sequences and accurately and intensively annotate them for gene content.

Figures and Tables

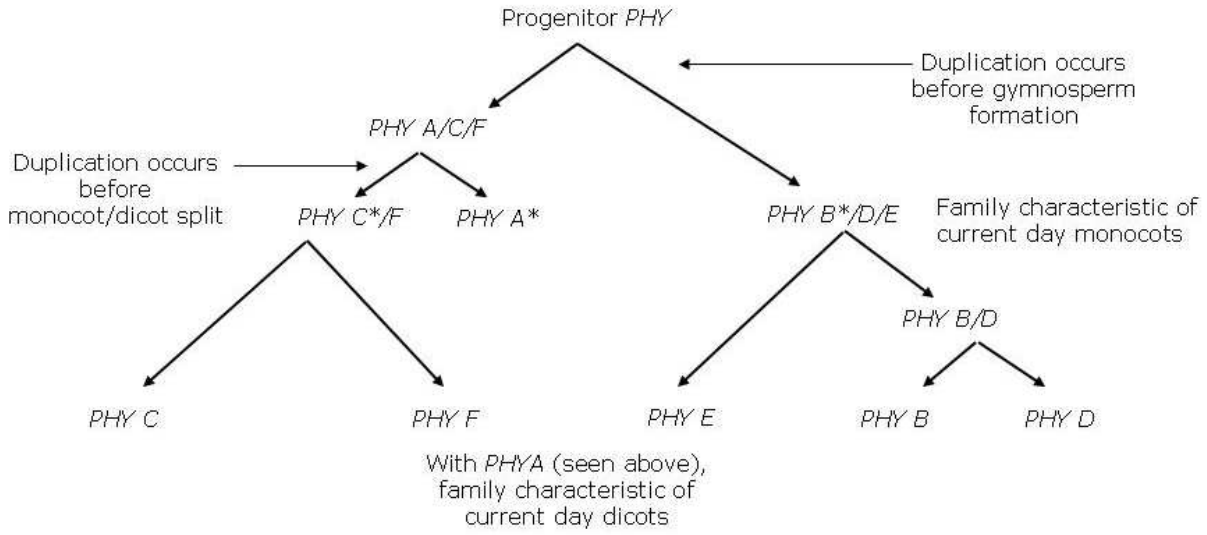


Figure 2.1: Overview of *PHY* evolution. A progenitor *PHY* presumably duplicated to yield the *PHYA/C/F* and *PHYB/D/E* subfamilies prior to gymnosperm formation. *PHYA/C/F* then underwent a second duplication to form the *PHYA* and *PHYC/F* subfamilies prior to monocot/dicot formation. *PHYC/F* and *PHYB/D/E* can then further duplicate into *PHYC*, *F*, *B*, *D*, and *E*, respectively, depending on the specific nature of a species' *PHY* family evolution.

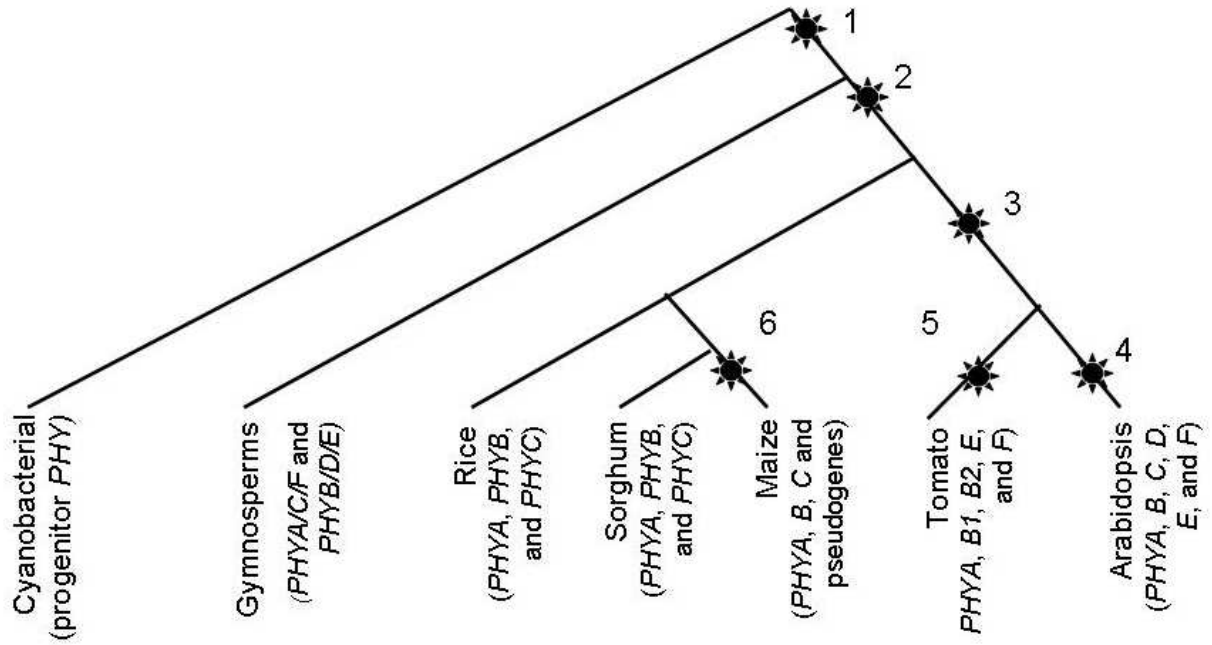


Figure 2.2: Phylogeny of *PHY* Evolution: (1) A progenitor *PHY* presumably duplicated to yield the *PHYA/C/F* and *PHYB/D/E* subfamilies prior to gymnosperm formation. (2) *PHYA/C/F* then underwent a second duplication to form the *PHYA* and *PHYC/F* subfamilies prior to monocot/dicot formation. This is the *PHY* family (*A, B, C*) characteristic of current day monocots such as rice. (3) and (4) *PHYC/F* and *PHYB/D/E* can then further duplicate into *PHYC, F, B, D,* and *E*, respectively, depending on the specific nature of a species' *PHY* family evolution. The five-membered *PHYA, B, C, D, E,* and *F* is characteristic of current day dicots such as Arabidopsis. There are numerous examples of species-independent evolution of the PHY family such as the presence of (5) a second member of the *PHYB/D* subfamily in tomato and (6) the presence of a *PHY* pseudodgenes in maize.

References

Ahmad, M., Jarillo, J.A., Cashmore, A.R. (1998) Chimeric proteins between cry1 and cry2 Arabidopsis blue light photoreceptors indicate overlapping functions and varying protein stability. *Plant Cell*., **10**: 197-207.

Ahmad, M., and Cashmore, A. R. (1996) The pef mutants of *Arabidopsis thaliana* define lesions early in the phytochrome signaling pathway. *Plant J*., **10**: 1103-1110.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**:403-410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**: 3389–3402.

Arber, E. A. N., and Parkin, J. (1907) On the origin of angiosperms. *Bot J. Linn. Soc*., **38**: 29-80.

Arumuganathan, K. and Earle, E.D. (1991) Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.,* **9**: 208-219.

Atchley, W. R. and Fitch, W. M. (1997) A natural classification of the basic helix-loop-helix class of transcription factors. *Proc. Natl. Acad. Sci. U.S.A.,* **94**: 5172-5176.

Aukerman M. J., Hirschfeld, M., Wester, L., Clack, T., Amasino, R. M., and Sharrock, R. A. (1997) A deletion in the PHYD gene of Arabidopsis Wassilewskija ecotype defines a role for phytochrome D in red/far-red light sensing. *Plant Cell,* **9**: 1317–1326.

Axelrod, D. I. (1952) A theory of angiosperm evolution. *Evolution*, **6**: 29-60.

Bancroft, I. (2000) Insights into the structural and functional evolution of plant genomes afforded by the nucleotide sequences of chromosomes 2 and 4 of *Arabidopsis thaliana*. *Yeast*, **17**: 1-5.

Bessey, C.E. (1915) The phylogenetic taxonomy of flowering plants. *Ann. Missouri Bot. Gard.*, **2**: 109-164.

Bews, J.W. (1927) Studies in the ecological evolution of angiosperms. *New Phytol*., **26**: 1-21.

Birney, E., Clamp, M., and Durbin, R. (2004) Genewise and genomewise. *Genome Res.*, **14**: 988-995.

Boguski, M., Lowe, T., Tolstoshev, C. (1993) dbEST—database for “expressed sequence tags”. *Nat Genet*., **4**: 332-333.

Borodovsky, M., and McInich, J. (1993) GENMARK: parallel gene recognition for both DNA strands. *Comput. Chem*. **17**: 123-133.

Borodovsky, M., Rudd, K.E., and Doonin, E.V. (1994) Intrinsic and extrinsic approaches for detecting genes in a bacterial genome. *Nucleic Acids Res*., **22**: 4756-4767.

Borthwick, H.A., Hendricks, S.B., Parker, M.W., Toole, V.K. (1952) A reversible photoreaction controlling seed germination. *Proc. Natl. Acad. Sci., USA,* **38**: 662-666.

Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol*. **268**: 78-94.

Butler, W., Norris, K. H., and Hendricks, S. B. (1959) Detection, assay, and preliminary purification of the pigment controlling photoresponsive development in plants. *Proc. Natl. Acad. Sci., USA*, **45**: 1703-1708.

Cadigan, K.M., Grossniklaus, U., Gehring, W.J. (1994) Functional redundancy: the respective roles of the 2 sloppy paired genes in Drosophila segmentation. *Proc. Natl. Acad., Sci. USA,* **91**: 6324-6328.

Chen, M., SanMiguel, P., De Oliveira, A.C., Woo, S.S., Zhang, H., Wing, R.A., Bennetzen, J.L. (1997) Microcolinearity in sh2-homologous regions of the maize, rice, and sorghum genomes. *Proc. Natl. Acad. Sci., USA*, **94**: 3431-3435.

Choi, G., Yi, H., Lee, J., Kwon, Y.-K., Soh, M.S., Shin, B., Luka, Z., Hahn, T.-R., Song, P.-S. (1999) Phytochrome signaling is mediated through nucleoside diphosphate kinase 2. *Nature*, **401**: 610-613.

Christensen, A.H. and Quail, P.H. (1989) Structure and expression of a maize phytochrome-encoding gene. *Gene* **85** (2): 381-390.

Clack, T., Matthews, S., Sharrock, R.A. (1994) The phytochrome apoprotein family in Arabidopsis is encoded by five genes: the sequences and expression of PHYD and PHYE. *Plant Mol. Biol*., **25**: 413-427.

Clamp, M., Andrews, D., Barker, D., Bevan, P., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Hubbard, T., Kasprzyk, A., Keefe, D., Lehvaslaiho, H., Iyer, V., Melsopp, C., Mongin, E., Pettett, R., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik. I,, Birney. E. (2003) Ensembl 2002. *Nucleic Acids Res*., **31**: 98-42.

Cronquist, A. (1988) The evolution and classification of flowering plants. The New York Botanical Garden, New York.

Curwen, V., Eyras, E, Andrews, D., Clarke, L., Mongin, E., Searle, S., Clamp, M. (2004) The Ensembl Automatic Gene Annotation System. *Genome Res*., **14**: 942-950.

Darwin, C. (1895) The power of movement in plants. New York: D Appleton and Co.

Davis, S. J., Kurepa, J., Vierstra, R. D. (1999) The *Arabidopsis thaliana* HY1 locus, required for phytochrome-chromophore biosynthesis, encodes a protein related to heme oxygenases. *Proc Natl Acad Sci, USA,* **96**: 6541–6546.

Dehal, P., Predki, P., Olsen, A.S., Kobayashi, A., Folta, P., Lucas, S., Land, M., Terry, A., Ecale, Zhou, C.L., Rash, S., Zhang, Q., Gordon, L., Kim, J., Elkin, C., Pollard, M.J., Richardson, P., Rokhsar, D., Uberbacher, E., Hawkins, T., Branscomb, E., Stubbs, L. (2001) Human chromosome 19 and related regions in mouse: conservative and lineage-specific evolution. *Science,* **293**: 104-111.

Dehesh, K., Franci, C., Parks, B. M., Seeley, K. A., Short, T. W., Tepperman, J. M., and Quail, P. H. (1993) Arabidopsis HY8 locus encodes phytochrome A. *Plant Cell,* **5**: 1081–1088.

Deng, X.W. and Quail , P. H. (1999) Signaling in light-controlled development. *Semin. Cell Dev. Biol*., **2**: 117-119.

Devlin, P.F., Robson, P.R.H., Patel, S.R., Goosey, L., Sharrock, R.A., and Whitelam, G.C. (1999). Phytochrome D acts in the shade-avoidance syndrome in Arabidopsis by controlling elongation and flowering time. *Plant Physiol*., **119** ; 909–915.

Devlin, P.F. (2002) Signs of the time: environmental input to the circadian clock. *J Exp Bot*., **53**: 1535-1550.

Devos, K.M., Gale, M.D. (1997) Comparative genetics in the grasses. *Plant Mol. Biol*., **35**: 3-15.

Devos, K.M., Beales, J., Nagatamura, Y., and Sasaki, T. (1999) Arabidopsis-Rice: will colinearity allow gene prediction across the Eudicot-Monocot divide? *Genome Res*., **9**: 825-829.

Devos, K.M, Gale, M.D (2000) Genome relationships: the grass model in current research. *Plant Cell*, **12**: 637-646.

Donoghue M.J., Mathews, S. (1999) Duplicate genes and the root of angiosperms, with an example using phytochrome sequences. *Mol Phylogenet Evol*., **9**: 489-500.

Donoghue (1994) Progress and prospects in reconstructing plant phylogeny. *Ann. Missouri Bot. Gard.* **81**: 405-418.

Dowell, R., Jokerst, R., Day, A., Eddy, S.R., Stein, L. (2001) The distributed annotation system. *BMC Bioinformatics*. **2**: 7.

Dunford, R.P., Kurata, N., Laurie, D.A., Money, T.A., Minobe, Y., and Moore, G. (1995) Conservation of fine-scale DNA marker order in the genomes of rice and the Triticeae. *Nucleic Acids Res*., **23**: 2724-2728.

Elich, T.D., and Chory, J. (1997) Phytochrome: if it looks and smells like a histidine kinase, is it a histidine kinase? *Cell,* **91**: 713-716.

Enright, J,T. (1982) Sleep movements of leaves: in defense of Darwin's interpretation. *Oecologia,* **54**: 253–259.

Fankhauser C., Yeh, K.C., Lagarias, J.C., Zhang, H., Elich, T.D., Chory, J. (1999) PKS1, a substrate phosphorylated by phytochrome that modulates light signaling in Arabidopsis**.** *Science*, **284**: 1539-1541.

Feuillet, C., Keller, B. (1999) High gene density is conserved at syntenic loci of small and large grass genomes. *Proc Natl. Acad. Sci*., *USA,* **96**: 8665-8670.

Flint, L. H. and E. D. McAlister (1935) Wavelengths of radiation in the visible spectrum inhibiting the germination of light-sensitive lettuce seed. *Smithson. Misc. Collect*., **94**: 1-11.

Florea, L., Hartzell, G. Zhang, Z., Rubin, G.M., Miller, W. (1998) A computer program for aligning cDNA sequence with a genomic DNA sequence. *Genome Res*., **8**: 967-974.

Franklin, K.A., Davis, S.J., Stoddart, W.M., Vierstra, R.D., Whitelam, G.C. (2003) Mutant analyses define multiple roles for phytochrome C in Arabidopsis photomorphogenesis. *Plant Cell*, **15**: 1981-1989.

Furuya, A. and Song, P. (1994) Assembly and properties of holophytochrome. In *Photomorphogenesis of Plants*, 2nd ed. Kluwer Academic Publishers, Dordrectht, The Netherlands, pp. 105-140.

Gale, M. and Devos, K. (1998) Comparative genetics in the grasses. *Proc. Natl. Acad. Sci., USA*, **95**: 1971-1974.

Garner, W. W. and Allard, H.A. (1933) Comparative responses of long-day and short-day plants to relative length of day and night. *Plant Physiol*., **8**: 347-356.

Gibson, T. and Spring, J. (1998) Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multiprotein domains. *Trends Genet*., **14**: 46-49.

Goodall, G., Filipowicz, W. (1989) The AU-rich sequences present in the introns of plant nuclear pre-mRNAs are required for splicing. *Cell*. **58**: 473-483.

Gray, M.W., Doolittle, W. F. (1982) Has the endosymbiont hypothesis been proven? *Microbiol. Rev*., **46**: 1-42.

Gu, Z. (2003). Role of duplicate genes in genetic robustness against null mutations. *Nature*, **421**: 63-66.

Harrison, P.M., Echols, N., and Gerstein, M.B. (2001) Digging for dead genes: an analysis of the characteristics of the pseudogene population in the *C. elegans* genome. *Nucleic Acids Res*., **29**: 818-830.

Harrison, P. and Gerstein, M. (2002a) Studying genomes through the aeons: protein families, pseudogenes, and proteome evolution. *J. Mol. Biol*., **318**:1155-1174.

Harrison, P., Kumar, A., Lan, N., Echols, N., Snyder, M., Gerstein, M. (2002b) A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution. J. Mol. Biol., 316: 409-419.

Harrison, P.M., Hegyi, H., Balasubramanian, S., Luscombe, N.M., Bertone, P., Echols, N., Johnson, T., Gerstein, M. (2002) Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Res.*, **12**: 272-280.

Harrison, P.M., Milburn, D., Zhang, Z., Bertone, P., Gerstein, M. (2003) Identification of pseudogenes in the *Drosophila melanogaster* genome. *Nucleic Acids Res*., **31**: 1033-1037.

Hauser, B.A., Cordonnier-Pratt, M.-M., Daniel-Vedele, F., Pratt, L.H. (1995) The phytochrome gene family in tomato includes a novel subfamily. *Plant Mol. Biol*., **29**: 1143-1155.

Hendricks, S. (1964) Photochemical aspects of plant photoperiodicity. *Photophysiology* **1**: 305-331.

Hennig, L., Funk, M., Whitelam, G.C., and Schäfer, E. (1999). Functional interaction of cryptochrome 1 and phytochrome D. *Plant J.*, **20**: 289–294.

Hirschfeld M, Tepperman, J.M, Clack, T., Quail, P.H., Sharrock, R.A. (1998) Coordination of phytochrome levels in phyB mutants of Arabidopsis as revealed by apoprotein-specific monoclonal antibodies. *Genetics*, **149**: 523–535.

Hoecker U., Xu, Y., Quail, P. H. (1998) SPA1: a new genetic locus involved in phytochrome A-specific signal transduction. *Plant Cell*, **10**: 19-33.

Howe, G.T., Bucciaglia, P.A., Hackett, W.P., Furnier, G.R., Cordonnier-Pratt, M.-M., Gardner, G. (1998) Evidence that the phytochrome gene family in black cottonwood has one *PHYA* locus and two *PHYB* loci but lacks members of the *PHYC/F* and *PHYE* subfamilies. *Mol. Biol. Evol.* **2**: 160-75.

Hudson, M., Ringli, C., Boylan, M. T., Quail, P. H. (1999) The FAR1 locus encodes a novel nuclear protein specific to phytochrome A signaling. *Genes Dev*., **13**: 2017-2027.

Johnson, E., Bradley, M., Harberd, N.P., and Whitelam, G.C. (1994). Photoresponses of light-grown phyA mutants of Arabidopsis. *Plant Physiol*., **105**: 141–149.

Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., Naruo, K., Okumura, S., Shimpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M., Tabata, S. (1996) Sequence analysis of the genome of the unicellular cyanobacterium Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res*., **3**: 109-136.

Keller, B., Feuillet, C. (2000) Colinearity and gene density in grass genomes. *Trends in Plant Sci.*, **5**: 246-251.

Kent, W.J., Zahler, AM. (2000) Conservation, regulation, synteny, and introns in a large-scale *C. briggisae-C. elegans* genomic alignment. *Genome Res*., **10**: 1115-1125.

Killian, A., Chen, J., Han, F., Steffenson, B., Kleinhoffs, A. (1997) Towards map-based cloning of barley stem rust resistance gene Rpg1 and rpg4 using rice as a intergenomic cloning vehicle. *Plant Mol. Biol*., **35**: 187-195.

Kircher, S., Kozma-Bognar, L., Adam, E., Harter, K., Schafer, E., Nagy, F. (1999) Light quality-dependent nuclear import of the plant photoreceptors phytochrome A and B. *Plant Cell,* **11**: 1445-1156.

Kneip, C., Hildebrandt, P., Schlamann, W., Braslavsky, S.E., Mark, F., Schaffner, K. (1999) Protonation state and structural changes of the tetrapyrrole chromophore during the Pr --> Pfr phototransformation of phytochrome: a resonance Raman spectroscopic study. *Biochemistry*, **38**: 15185-15192.

Kolukisaoglu, H. Ü., Braun, B., Martin, W.F., Schnieder-Poetsch, H.A.W. (1993) Mosses do express conventional, distantly B-type related phytochromes: Phytochrome of *Physomitrella patens*. *Hedw. Fed. Eur. Biochem. Soc. Lett*. **334**:95-100.

Koornneef, M., Rolff, E., and Spruit, C.J.P. (1980) Genetic control of light-inhibited hypocotyl elongation in *Arabidopsis thaliana* (L.) *Heynh. Z Pflanzenphysiol*, **100**: 147–160.

Kuno, N., Furuya, M. (2000) Phytochrome regulation of nuclear gene expression in plants. *Semin Cell Dev Biol*., **11**:485-493.

La Rota, M., and Sorrells, M. E. (2004) Comparative DNA sequence analysis of mapped wheat ESTs reveals the complexity of genome relationships between rice and wheat. *Funct Integr Genomics*. 2004 **4**: 34-46.

Lamparter, T., Mattmann, F., Gärtner, W., Börner, T., Hartmann, E., Hughes, J. (1997) Characterization of recombinant phytochrome from the cyanobacterium Synechocystis. *Proc. Natl. Acad. Sci., USA*, **94**: 11792-11797.

Lagarias, J. D. (1985) Progress in the molecular analysis of phytochrome. *Photochem. Photobiol*., 42: 811-820.

Li, W.Z. and Chinappa, C.C. (2004) Isolation and characterization of PHYC gene from *Stellaria longipes*: differential expression regulated by different red/far-red light ratios and photoperiods. *Planta.* July29 (ePub ahead of print; PMID: 15290294.

Liu, Y., Harrison, P.M., Kunin, V., Gerstein, M. (2004) Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of horizontally transferred genes. *Genome Biology*, **5**: R64.1-R64.11.

Lin, C., Yang, H., Guo, H., Mockler, T., Chen, J., Cashmore, A.R. (1998) Enhancement of blue-light sensitivity of Arabidopsis seedlings by a blue light receptor cryptochrome 2. *Proc. Natl. Acad. Sci. USA*, **95**: 2686-2690.

Littlewood T.and G. I. Evan (1998) *Helix-Loop-Helix Transcription Factors, 3rd edition*. Oxford Univ. Press, New York, New York.

Lynch, M. and Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Cur. Opin. Microbiol*., **2**: 548-554.

Lynch, M., O'Hely, M., Walsh, B., Force, A. (2001) The probability of preservation of a newly arisen gene duplicate. *Genetics*, **159**: 1789-1804.

Mathé, M., Sago, M.-F., Schiex, T., and Rouzé, P. (2002) Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res*., **30**: 4103-4117.

Mathews, S. and Sharrock, R.A. (1996) The Phytochrome Gene Family in Grasses (Poaceae): A Phylogeny and Evidence that Grasses Have a Subset of the Loci Found in Dicot Angiosperms. *Mol. Biol. Evol*. **13**: 1141-1150.

Mathews, S., Lavin, M., Sharrock, R.A. (1995) Evolution of the phytochrome gene family and its utility for phylogenetic analysis of angiosperms. *Ann. Mo. Bot. Gard*., **82**: 266-321.

Mathews, S., Lavin, M., Sharrock, R.A. (1996) The phytochrome gene family in grasses (Poacaea): a phylogeny and evidence that grasses have a subset of the loci found it dicot angiosperms. *Mol. Biol. Evol*., **13**: 1141-1150.

Mathews, S., Burleigh, J.G., and Donoghue, M.J. (2003) Adaptive Evolution in the Photosensory Domain of Phytochrome A in Early Angiosperms. *Mol. Biol. Evol*., **20**: 1087-1097.

Mathews, S., Sharrock, R.A. (1997) Phytochrome gene diversity. *Plant Cell Environ.*, **20**: 666–671.

Mayer, K., Murphy, G., Tarchini, R., Wambutt, R., Volckaeart, G., Pohl, T., Dusterrhoft, A., Stiekema, W., Entian, K.D., Terryn, N., Lemcke, K., Haase, K., Hall, C.R., Van Dodeweerd, A.M., Tingey, S.V., Mewes, H.W., Bevan, M.W., Bancroft, I. (2001) Conservation of microstructure between a sequenced region of rice and multiple segments of the genome of *Arabidopsis thaliana*. *Genome Res*., **11**: 1167-1174.

Meischke, D. (1936) Über den Einfluss der Strahlung aug Licht- und Dunkelkeimer. *Jahrb. Wiss. Bot*. **83**: 359-405.

Mighell, A.J., Smith, N.R., Robinson, P.A., Markham, A.F. (2000) Vertebrate pseudogenes. *FEBS Lett*., **468**: 109-114.

Millar, A.J., Kay, S.A. (1991) Circadian control of cab gene transcription and mRNA accumulation in Arabidopsis. *The Plant Cell*, **3**: 541–550.

Morgenstern, B. (2000) A space-efficient algorithm for aligning large genomic sequences. *Bioinformatics*, **16**: 948-949.

Muramoto, T., Kohchi, T., Yokota, A., Hwang, I., Goodman, H. M. (1999) The Arabidopsis photomorphogenic mutant hy1 is deficient in phytochrome chromophore biosynthesis as a result of a mutation in a plastid heme oxygenase. *Plant Cell*, **11**: 335–348.

Nadeau, J., and Sankoff, D. (1997) Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. *Genetics*, **14**7: 1259-1266.

Nagatani, A., Reed, J. W. and Chory, J. (1993) Isolation and initial characterization of Arabidopsis mutants that are deficient in phytochrome A. *Plant Physiol.*, **102**: 269–277.

Neff, M.M., and Chory, J. (1998). Genetic interaction between phytochrome A, phytochrome B and cryptochrome 1 during Arabidopsis development. *Plant Physiol.*, **118**: 27–36.

Ni, M., Tepperman, J. M., Quail, P. H. (1998) PIF3, a phytochrome interacting factor necessary for normal photoinduced signal transduction is a novel basic helix-loop-helix protein. *Cell*, **95**: 657-667.

Ni, M., Tepperman, J., Quail, P. (1999) Binding of Phytochrome B to its nuclear signaling partner PIF3 is reversibly induced by light. *Nature,* **400**: 781-784.

Nowak, M. A., Boerlijst, M.C., Cooke, J., Smith, J.M. (1997) Evolution of genetic redundancy. *Nature*, **388**: 167-171.

Parks, B. M. and P. H. Quail (1993) hy8, a new class of Arabidopsis long hypocotyl mutants deficient in functional phytochrome A. *Plant Cell*, **5**: 39–48.

Pasentsis, K., Paulo, N., Algarra, P., Dittrich, P., Thümmler, F. (1998) Characterization and expression of the phytochrome gene family in the moss *Ceratodon purpureus*. *Plant Journal* **13**: 51-61.

Paterson, A.H., Lin, Y.R., Schertz, K.F., Doebley, J.F, Pinson, S.R.M., Liu, S.C., Stansel, J.W., Irving, J.E. (1995) Convergent domestication of cereal crops by independent mutations at corresponding genetic loci. *Science*, **269**: 1714-1718.

Paterson, A.H., Lan, T.H., Reischmann, K.P., Chang, C., Lin, Y.R., Liu, S.C., Burow, M.D., Kowlaski, S.P., Katsar, C.S., Delmonte, T.A. (1996) Towards a unified genetic map, transcending the monocot-dicot divergence. *Nature Genet*., **14**: 380-382.

Pearson, W.R., and Limpman, D.J. (1998) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci., USA*, **85**: 2444-2448.

Pereira, M.G., Lee, M. (1995) Identification of genomic regions affecting plant height in sorghum and maize. *Theorh. Appl. Genet*., **90**: 380-388.

Pratt, L.H., Cordonnier-Pratt, M.-M., Hauser, B., Caboche, M. (1995) Tomato contains two differentially expressed genes encoding B-type phytochromes, neither of which can be considered an ortholog of Arabidopsis phytochrome B. *Planta*. **197**(1):203-6.

Pratt, L.H. (1982) Phytochrome: The protein moiety. *Annu. Rev. Plant Physiol*., **33**: 557-582.

Pratt, L. H., Cordonnier-Pratt, M.-M., Kelmenson, P. M., Lazarova, G., I., Kubota, T., and Alba, R.M. (1997) The phytochrome gene family in tomato. *Plant Cell Environ*., **20**: 672-677.

Quail, P.H. (1991) Phytochrome: A light-activated molecular switch that regulates plant gene expression. *Annu. Rev. Genet*., **25**: 389-409.

Quail, P.H., (1997). An emerging molecular map of the phytochromes. *Plant Cell Environ.,* **20**: 657-665.

Quail, P. H. (1998) The phytochrome family: dissection of functional roles and signaling pathways among family members. *Philos. Trans. R. Soc. London,* **353**: 1399-1403.

Quail, P.H. (2002). Phytochrome photosensory signaling networks. *Nat. Rev. Mol. Cell Biol*., **3**: 85–93

Reed, J. W., Nagatani, A., Elish, T.D., Fagan, M., Chory, J. (1993) Phytochrome A and phytochrome B have overlapping but distinct functional in Arabidopsis development. *Plant Physiol*., **104**: 1139-1149.

Reed, J. W., Nagpal, F., Poole, D. S., Furuya, M., Chory, J. (1994) Mutations in the gene for the red/far-red light receptor phytochrome B alter cell elongation and physiological responses throughout Arabidopsis development. *Plant Cell*, **5**: 147-157.

Rouzé, P., Pavy, N., and Rombauts, S. (1999) Genome annotation: which tools do we have for it? *Curr. Opin. Plant Bio*., **2**: 90-95.

Saga, Y., Yagi, T., Ikawa, Y., Sakakura, T., Aizawa, S. (1992). Mice develop normally without tenascin. *Genes Dev*. **19**: 1821-1831.

Sakata, K., Nagamura, Y., Numa, H., Antonio, B.A., Nagasaki, H., Idonuma, A., Watanabe, W., Shimizu, Y., Horiuchi, I., Matsumoto, T., Sasaki, T., Higo, K. (2002) RiceGAAS: an automated system for rice genome annotation. *Nucleic Acids Res*. **30**: 98-102.

Salse, J. Piegu, B., Cooke, R., Delseny M. (2002) Synteny between *Arabidopsis thaliana* and rice at the genome level: a tool to identity conservation in the ongoing rice genome sequencing project. *Nuc Acids Res*., **30**: 2316-2328

Sanderson, M.J., and Doyle, J.A. (2001) Sources of error and confidence intervals in estimating the age of angiosperms from rbcL and 18S rDNA data. *Am J. Bot*., **88**: 1499-1516.

Sato, N. (1990) Nucleotide sequence of a pseudogene for pea phytochrome reminiscent of an incorrect spelling event. *Nucl. Acid Res*. **18** (12): 3632.

Schneider-Poetsch, H. A. W., Kolukisaoglu, U., Clapham, D.H., Hughes, J., Lamparter, T. (1998) Non-angiosperm phytochromes and the evolution of vascular plants. *Physiol. Plantarum*, **102**: 612-622.

Sharrock, R. A., and Quail, P.H. (1989) Novel phytochrome sequences in *Arabidopsis thaliana*: structure, evolution, and differential expression of a plant regulatory photoreceptor family. *Genes Dev.* **3**: 1745-1757.

Shinomura, T., Nagatani, A., Hanzawa, H., Kubota, M., Watanabe, M., Furuya, M. (1996) Action spectra for phytochrome A- and B-specific photoinduction of seed germination in *Arabidopsis thaliana*. *Proc Natl Acad Sci, USA,* **93**: 8129–8133.

Smith, H., Whitelam, G. C., and McCormac, A. C. (1991) Do the members of the phytochrome family have different roles? Physiological evidence from wildtype, mutant and transgenic plants. B Thomas and B C Jhonson, Editors, In: *Phytochrome Properties and Biological Action*, Springer-Verlag, Berlin, pp. 217–236.

Smith, H., and Whitelam, G.C. (1997). The shade avoidance syndrome: Multiple responses mediated by multiple phytochromes. *Plant Cell Environ.*, **20**: 840–844.

Soh, M. S., Hong, S. H. Hanzawa, H. Furuya, M. Nam, H. G. (1998) Genetic identification of FIN2, a far red light-specific signaling component of Arabidopsis thaliana. *Plant J*., **16**: 411-419.

Somers, D.E., Sharrock, R.A., Tepperman, J.M., and Quail, P.H. (1991). The hy3 long hypocotyl mutant of Arabidopsis is deficient in phytochrome B. *Plant Cell,* **3**: 1263–1274.

Somers, D.E., Devlin, P.F., Kay, S.A. (1998) Phytochromes and cryptochromes in the entrainment of the Arabidopsis circadian clock. *Science,* **282**: 1488–1490.

Takano, M. Kanegae, H., Shinomura, T., Miyao, A., Horochika, H., Furuya, M. (2001) Isolation and characterization of rice phytochrome A mutants. *Plant Cell,* **13**: 521-534.

Takhtajan, A. L. (1969) Flowering plants: origin and dispersal. Translated by C. Jeffrey. *Smithsonian Institution Press*, Washington, D.C., pp. 310.

Tandeau de Marsac, N., Hournard, J. (1993) Adaptation of cyanobacteria to environmental stimuli: new steps towards molecular mechanisms. *FEMS Microbiol. Rev*., **104**: 119-190.

Tarchini, R., Biddle, P., Wineland, R., Tingey, S., Rafalski, A. (2000) The complete sequence of 340 kb of DNA around the rice Adh1-Adh2 region reveals interrupted colinearity with maize chromosome 4. *Plant Cell,* **12**: 381-391.

Thorne, R. R. (1974) A phylogenetic classification of the Anoniflorae. *Aliso*., **8**: 147-209.

Tikhonov, A.P., SanMiguel, P.J., Nakajima, Y., Gorenstein, N.M., Bennetzen, J.L., Avramova, Z. (1999) Colinearity and its exceptions in orthologous adh regions of maize and sorghum. *Proc Natl Acad Sci U S A*, **96**: 7409-7414.

Torrents, D., Suyama, M., Zdobnov, E., Bork, P. (2003) A genome-wide survey of human pseudogenes. *Genome Res*., **13**: 2559-2567.

Van Dodeweerd, A.M., Hall, C.R., Bent, E.G., Johnson, S.J., Bevan, M.W., and Bancroft, I. (1999) Identification and analysis of homeologous segments of the genomes of rice and *Arabidopsis thaliana*. *Genome*, **42**: 887-892.

Vanin, E.F. (1985) Processed pseudogenes: characteristics and evolution. *Annu. Rev. Genet*., **19**: 253-272.

Vision, T.J, Brown, D.G., and Tanksley, S.D. (2000) The origins of genomic duplications in Arabidopsis. *Science*, **290**: 2114-2117.

Wagner D., Hoecker, U., Quail, P. H. (1997) RED1 is necessary for phytochrome B-mediated red light-specific signal transduction in Arabidopsis. *Plant Cell*, **9**: 731-743.

Wagner, A. (2000) Robustness against mutations in genetic networks in yeast. *Nature Genet*. **31**:400-404.

Wang, Z.M., LeTheirry d'Ennequin, M., Panaud, M., Gale, M.D., Sarr, A., Devos, K.M. (2001) Trait mapping in foxtail millet. *Theor. Appl. Genet*. In press

Ware, D.H., Jaiswal, P., Ni, J., Yap, I.V., Pan, X., Clark, K.Y., Teytelman, L., Schmidt, S.C., Zhao, W., Chang, K., Cartinhour, S., Stein, L.D., McCouch, S.R. (2002) Gramene, a tool for grass genomics. *Plant Physiol*. **130**: 1606-1613.

Whellan, S.J., Church, D.M., and Ostell, J.M. (2001) Spidey: a tool for mRnA-to-genomic alignments. *Genome Res*. **8**: 967-974.

Whitelam, G. C., Johnson, E., Peng, J., Carol, P., Anderson, M. L., Cowl, J. S., Harberd, N. P. (1993) Phytochrome A null mutants of Arabidopsis display a wild-type phenotype in white light. *Plant Cell* **5**: 757–768.

Whitelam, G.C., Patel, S., Devlin, P.F. (1998) Phytochromes and photomorphogenesis in Arabidopsis. *Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences*, **353**: 1445–1453.

Winzeler, E., Shoemaker, D.D., Astromoff, a., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito R., Boeke, J.d., Bussey H., Chu, A.M., Connelly, C., Davis, K., Dietrich, F., Dow, S.W., El Bakkoury, M., Foury, F., Friend, S.H., Gentalen, E., Giaever, G., Hegemann, J.H., Jones, T.,

Laub, M., Liao, H., Davis, R.W. (1999) Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science,* **285**: 901-906.

Wolfe, K., Guoy, M., Yang, Y., Sharp, P., Li, W.H. (1989) Date of monocot/dicot divergence estimated from chloroplast DNA sequence data. *Proc. Natl. Acad. Sci., USA,* **86**: 5201-5202.

Wu, C.H., Yeh. L.-S. L., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., Hu, H.-H., Ledley, R., Kouresis, P., Suzek, B., Vinayake, C.R., Zhang, J., Barker, W. (2003) The Protein Information Resource. *Nucl. Acids. Res*. **31**: 345-347.

Yamaguchi, R., Nakamura, M., Mochizuki, N., Kay, S. A., Nagatani, A. J. (1999) Light-dependent translocation of a phytochrome B-GFP fusion protein to the nucleus in transgenic Arabidopsis. *Cell Biol.,* **145**: 437-445.

Yan, L., Loukoianov, A., Tranquilli, G. Helguera, M., Fahima, T. and Dubcovsky, J. (2003) Positional cloning of wheat vernalization gene VRN1. *Proc. Natl. Acad. Sci., USA,* **100**: 6263-6268.

Yanovsky, M.J., Casal, J.J., Whitelam, G.C. (1995) Phytochrome A, phytochrome B and HY4 are involved in hypocotyl growth responses to natural radiation in Arabidopsis: weak de-etiolation of the phyA mutant under dense canopies. *Plant Cell Environ,* **18**: 788–794.

Yeh, K.-C., Wu, S.-H., Murphy, J.T., Lagarias, J.C. (1997) A cyanobacterial phytochrome two-component light sensory system. *Science,* **277**: 1505-1508.

Yeh, K.C., Lagarias, J.C. (1998) Eukaryotic phytochromes: light-regulated serine/threonine protein kinases with histidine kinase ancestry. *Proc Natl Acad Sci, U S A,* **95**: 13976-81.

Zhang, Z., Harrison, P., Gerstein, M. (2002) Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. *Genome Res*., **12**: 1466-1482.

CHAPTER 3

DEVELOPMENT OF A SYSTEM FOR DETAILED ANNOTATION

Introduction

Numerous tools currently exist to facilitate the large-scale annotation of genome sequences. Vast amounts of data and several algorithms and computer programs are typically combined to increase the accuracy of gene prediction and annotation within a genome. As the annotation of whole genomes and genome sequences is not trivial (Rouzé *et al*., 1999), many database projects have arisen from the need both to store the large amounts of information generated through annotation efforts and to increase the speed by which genome sequences are annotated (Dowell *et al*., 2001; Sakata *et al*., 2002; Ware *et al*., 2002; Clamp *et al*., 2003; Birney *et al*., 2004). Many of the efforts within the genomic community thus far have focused on the large-scale annotation of whole genomes and the development of automatic annotation pipelines (for examples see Sakata *et al*., 2002; Ware *et al*., 2002; Clamp *et al*., 2003; Birney *et al*., 2004). This automated approach typically imparts a baseline annotation that allows manual curators to focus on the more difficult regions that cannot be handled with the automatic pipeline (Schoof and Karlowski, 2003). These databases and annotation systems are tremendously valuable to the genome community at-large as they provide massive amounts of data and ever-improving computer programs and algorithms for different components of the annotation process. Yet, due to the tremendous scope of the goal of whole genome annotation/database projects, the systems themselves are largely inaccessible to smaller laboratories with reduced needs and their use is extremely complex. Moreover, the implementation of these systems requires rather extensive computer hardware as well as advanced computer skills.

A new approach, described here, has been developed to facilitate the organized collection and display of all evidence needed for a detailed manual annotation across the typically smaller genomic regions studied in medium- to low-throughput sequencing projects. Subsequent to

annotation, informed and rational decisions can then be made about the identity of the appropriate gene model(s) to use for identifying genes. This method of annotation consists of a collection of scripts to help the non-computer scientist perform quickly, easily, and reliably a series of tasks that ultimately provide the user with a comprehensive view of all experimental and predicted evidence aligned under the sequence to be annotated, such that decision making by the scientist is greatly facilitated. This approach was designed to satisfy four overall goals: (1) combine *ab initio* gene prediction programs and sequence homology searches to nucleotide or amino acid databases to improve gene finding and to determine a final putative gene model, (2) develop an annotation system that allows multiple lines of evidence to be viewed interactively, such that the evidence can be curated by the viewer to facilitate improved annotation, (3) ensure that the method is flexible enough that low- to medium-throughput sequencing laboratories investigating a specific question regarding a specific region of a genome can benefit from this method, and (4) incorporate a database that can store the results from the gene prediction and sequence comparison analyses for subsequent query and/or retrieval.

The overall structure of this approach is most similar to GeneMachine (Makalowska *et al*., 2001) and Genescript (Hudek *et al*., 2003), two programs that rely on a series of gene prediction programs, similarity searches and perl modules to aid in annotation. GeneMachine, Genescript, and the approach described in the following all subscribe to the idea that the combination of homology-based and *ab initio* gene prediction approaches ultimately lead to more accurate gene models.

GeneMachine utilizes the gene prediction programs Genscan (Burge and Karlin, 1997) and FGenes (Soloyev *et al*., 1995) to determine the locations of potential genes within a DNA sequence. Basic local alignment search tool (BLAST; Altschul *et al*., 1990, 1997) is used to

compare the genome sequence to databases of known gene and protein sequences and the results are incorporated into the annotation. The user is allowed to define stringency thresholds for the return of sequence homology data for a more customized sequence homology search. The resulting data is processed by a series of perl modules that parse and write the outputs in ASN.1 format. Sequin (Benson *et al*., 1998), developed at the National Center for Biotechnology Information (NCBI), can then be used to visualize the outputs. Sequin, however, is not a particularly interactive annotation viewer, nor does it offer the extensive flexibility that is available through other annotation viewers such as Artemis (Rutherford *et al*., 2000) and Apollo (Lewis *et al*., 2002).

Genescript uses the gene prediction programs Genscan (Burge and Karlin, 1997) and HMMGene (Krogh, 1997) to predict potential coding regions that are then compared against the non-redundant (nr) protein database at NCBI. Searches are performed with BLAST to the expressed sequence tag (EST) database at NCBI and then refined with Sim4 (Florea *et al*., 1998) to identify intron-exon boundaries more accurately. Scores reflecting the likelihood of the existence of a gene are then compiled based on the gene prediction evidence and sequence similarity search results. The results are available in HTML, PDF, PNG, GenBank, and EMBL formats, the last of which can be imported into Artemis (Rutherford *et al*., 2000), a free genome viewer and annotation tool. For users who desire an interactive annotation viewer, Artemis allows visualization of sequence features within the context of the sequence and its six-frame translation.

The approach described here provides a number of advantages over Genescript and GeneMachine. These advantages include (i) the use of a relational database to store both gene prediction and BLAST data for subsequent query and retrieval, (ii) the ability to easily add

partial sequences of interest derived from the overall genomic sequence to refine the annotation through more intense investigation, and (iii) the possibility to work effectively with any gene prediction program provided that information in the output files of the program can be parsed and inserted into the database.

Description of Method

Once an accurate and complete genomic sequence is obtained, several steps lead to the final curated annotation. Initially, gene prediction programs are used to identify genomic regions likely to encode proteins. Subsequently, all evidence garnered via BLAST and other sequence comparison tools are accumulated and displayed against the full genomic sequence. These types of evidence include identity or strong similarity to ESTs, full- and partial-length cDNAs, proteins, known functional amino acid domains, and so on. The accumulated evidence is then critically compared to the gene prediction evidence and any overlapping evidence is analyzed. Thus, the final annotation can be assigned after careful manual curation.

It is the accumulation and evaluation of this evidence that the approach described here facilitates. Requisite scripts and database queries were designed to create output files that are compatible with Artemis (Rutherford et al., 2000). Thus, Artemis can be used as an electronic notebook with which to make final annotation decisions based on the available data displayed in an interactive graphical format.

Artemis was chosen as the visualization tool on which to base this method for four main reasons. First, Artemis was one of the few freely available annotation viewers in existence at the inception of this project that did not require programming knowledge to install and use. Second, the results of individual analysis programs such as gene predictors and blast results can be reformatted, imported into Artemis, and overlaid onto the sequence for visualization. Third, the

descriptors, or qualifiers, attached to each feature can be edited. For example, a feature in Artemis might be "blastn_hit" with "score" and "E-value" as qualifiers. This imparts flexibility to the annotator by allowing each annotated feature to be customized by modifying the attached qualifier. The user can easily move, delete, and modify any of the experimental evidence gathered in the analysis, thereby imparting a control over the final annotation that is not possible with the highly automated, yet less flexible systems. Fourth, Artemis can read and write nucleotide and amino acid sequences in multiple file formats, as well as read entries from many public databases. Sequences or annotation files can with Artemis be easily converted between multiple file formats, such as EMBL, GENBANK, and general feature format (GFF). The files can then be incorporated into other visualization programs or used by other algorithms for further analysis.

The generation of all of this information and the resulting capacity to view it in Artemis is accomplished by completing four main tasks: (1) predict potential gene locations and exon/intron boundaries (Figure 3.1A), (2) prepare all sequences to be analyzed further (Figure 3.1B), and (3) compare the sequences to a database or sequence repository, such as GenBank (Benson et al., 2003), for identification (Figure 3.1C), and (4) format all of this information in such a way that it can be viewed, analyzed, and edited in Artemis to produce a final and highly curated annotation. All of this data is stored within two database schemas: a gene prediction and analysis schema (Figure 3.1A) and a sequence comparison schema (Figure 3.1B).

Prediction of Probable Gene Locations and Exon/intron Boundaries (Figure 3.1A): Four gene prediction programs, and a second version of one of them (Table 3.1), are currently supported. After running any number or combination of the prediction programs on a sequence of interest (Figure 3.1A.i) the predictions must be reformatted in such a way that they can be visualized and

compared. These prediction programs produce output files of different formats. A perl CGI interface (Figure 3.2), activated by the perl script artparse.pl[1], performs two functions. First, the information requested within the interface prompts the user to provide the data necessary to track both the prediction programs employed and the genomic sequence being analyzed. For example, the *filename* prompt (Figure 3.2, first row of the interface) demands a standardized naming convention defined parenthetically to the right of the window. This information satisfies the uniqueness constraints of the database so that either the sequence being analyzed or the prediction results can be updated within the database without data loss or overlap occurring. For example, when an updated assembly file, or ace file as it is referred to in consed (Gordon *et al*., 1998), is generated, the prediction programs can simply be rerun on the updated sequence and resubmitted to the database. Second, upon completing the interface form, the execution of the *RUN* command uploads the prediction program outputs to the gene prediction and analysis schema[2] in an Oracle 9i relational database (Figure 3.1A.ii).

This execution of *RUN* from the artparse.pl interface uploads the gene prediction outputs to the database by parsing and reformatting them into a format consistent with the requirements of the schema. *RUN* also activates four additional perl scripts[3], one for each prediction program. The two versions of GeneMark described in Table 3.1 can be parsed by one script as their output files are identical. The gene prediction and analysis schema consists of one table per prediction program that stores all information produced by each individual program, a table dedicated to the storage of template information entered through the perl CGI interface, and a composite table

[1] The script artparse.pl was designed and written by A. Eastman and D. Kolychev.
[2] The gene prediction and analysis database schema was designed and implemented by S. Khosla, A. Eastman, and M. Shah.
[3] The perl scripts were designed and written by A. Eastman and D. Kolychev. The script names are: FGeneSH2Ora.pl, GeneMark2Ora.pl, Genscan2Ora.pl, and RiceHMM2Ora.pl.

that holds the start and stop locations of all predicted exons (Figure 3.3). This composite table permits, if desired, simple queries to the database to determine the number and location of regions with overlapping predictions.

A second perl script entitled getart.pl[4] queries the database to retrieve gene prediction data and generates a file in EMBL format for each prediction program output to be read into Artemis (Figure 3.1A.iii) for visualization and comparison (Figure 3.4). This graphical visualization serves multiple purposes. An overall consensus of multiple prediction programs can be easily created, and specific regions of the genomic sequence whose annotation might benefit from targeted homology searches to databases of known protein and amino acid sequences can be identified easily.

As stated in the introduction, one purpose of developing this method was the need for flexibility. For example, upon the publication of a review of plant-specific gene predictors, it was determined that a second, hidden markov model version of GeneMark needed to be incorporated in addition to the original version (Pavy *et al*., 1999). GeneMark.hmm was easily incorporated into this approach without data loss or overlap occurring due to the unique program IDs assigned in the database. Other prediction programs can be added to this analysis to enhance the annotation relatively easily. If their output forms differ from those programs listed in Table 3.1, however, modifications to perl scripts artparse.pl and getart.pl might be necessary. In addition to the five gene prediction programs specifically supported here, parsers for four other widely used gene prediction programs, Sim4 (Florea *et al*., 1998), Grail (Mural *et al*., 1992), ESTScan (Iseli, 1999) and MZEF (Zhang, 1997), are available in bioperl (bioperl.org). These,

[4] The perl script getart.pl was designed and written by A. Eastman and D. Kolychev.

and other, publicly available parsers can aid a user wishing to adapt this method with prediction programs other than those described here.

Preparation of All Sequences for Further Analysis (Figure 3.1B): Commands from the file menus within Artemis can be used to create many combinations of nucleotide or amino acid sequence files for further analysis (Figure 3.1B.i). As Artemis was not specifically designed to support exporting these multi-fastA files *en masse* to a database for sequence comparisons or analysis pipelines, modifications have to be made to the files before submitting them for analysis.

For example, a multi-fastA file was created containing the nucleotide sequences of the predicted exons from a particular gene prediction program by selecting those exons of interest and, from the Artemis menu, using the "Write/bases of selection/fastA format" command to create a file of multiple sequences in fastA format. The headers of each fastA sequence in this file contained the following information: the prediction program used to predict the exons, the start and stop position of the exon in the genome sequence, the strand direction (forward or reverse), and the predicted molecular weight of the polypeptide generated from the sequence. These headers do not contain any information about the name or version of the original genome sequence from which this subset of fastA sequences is generated, or the date of the generation of the fastA sequences. Therefore, these headers must be replaced with headers that better reflect the identity of these sequences. This same problem exists for sequence files of any type generated by Artemis.

The perl script header.pl[5] remedies this problem by renaming the fastA sequence headers in each user-generated multi-fastA file according to a standard naming convention that includes all of the relevant characteristics of each individual fastA sequence and satisfies the uniqueness constraints of the database (Figure 3.1B.ii). This script utilizes information from the original header of each sequence as well as the name of the file containing the multiple fastA sequences for which the headers need to be renamed. To rename the headers properly, the file containing the sequences must be named according to a pre-defined format so that the perl script is provided with all information necessary to complete the renaming process. Currently, header.pl can appropriately handle the following types of multi-sequence files generated by the user if the standard naming convention of the multi-fastA file is satisfied: nucleotide or amino acid sequences of the predicted genes, nucleotide or amino acid sequences of any open reading frame (ORF) in the genomic sequence greater than some number of amino acids as defined by the user, the genome sequence as a whole, and any other specific region of the genome sequence the researcher is interested in that can be extracted from the entire sequence using Artemis commands. Table 3.2 defines these types of files, the Artemis commands necessary to generate them, and the naming convention necessary for each type of multi-sequence file.

In addition to the following files generated from Artemis. The *BREAKUP* command from the Wisconsin Package of GCG (Genetics Computer Group; Womble, 2000) can be used to generate overlapping nucleotide sequences of some defined length from the genome sequence. A perl script entitled pf.pl[6] renames the headers of this multi-fastA file as described in Table 3.2 to satisfy database uniqueness constraints.

[5] The perl script header.pl was designed and written by A. Eastman and D. Kolychev.

[6] The perl script pf.pl was designed and written by A. Eastman and D. Kolychev.

Sequence Comparison and Identification (Figure 3.1C): After generating the sequences, comparisons can be made to databases of known nucleotide and protein sequences (Figure 3.1C.i). Multiple databases and comparison algorithms exist for such analyses and the files of multi-fastA sequences with renamed headers generated from Artemis or GCG can be compared to the many databases of publicly available sequences. These comparisons can generally be performed by any method preferred by the user. For example, high-throughput BLAST to existing NCBI databases or to a database prepared by the user (perhaps with the *FORMATDB* command in GCG) can be performed in a UNIX environment. Whatever the method of comparison, the BLAST output files containing the comparison results are then parsed and uploaded to the sequence comparison schema (Figure 3.1C) with the perl script blastToMagic_main.pl[7] (Figure 3.1C.ii), which employs existing BioPerl modules (bioperl.org).

The "sequence comparison schema" (Figure 3.5) described here is actually a previously existing portion of MAGIC_Annotation, a subschema of the Modular Approach to a Genomic Integrated and Comprehensive (MAGIC) database (Cordonnier-Pratt *et al*., 2004) developed at the Laboratory for Genomics and Bioinformatics at the University of Georgia. This portion of the MAGIC_Annotation subschema stores all data necessary to track multiple comparisons of a sequence to various databases. Implementation of the remainder of MAGIC_Annotation or the MAGIC database is not necessary to utilize the method described here. Hence, these tables are referred to here as the "sequence comparison schema" rather than MAGIC_Annotation.

A final perl script, artfromblast.pl[8] enables the user to query the sequence comparison schema for all comparison results above a user-defined score or below an E-value threshold and return the appropriate hits in an Artemis compatible format to visualize the results together with

[7]BlastToMagic_main.pl was written by F. Sun.
[8] artfromblast.pl was designed and written by A. Eastman and D. Kolychev.

the gene prediction data (Figure 3.1C.iii). This script is comprised of two portions: the executable script that parses the sequence comparison output file into a format read by Artemis, and a configuration file that allows the user to set parameters and cutoff values for the types of comparison results displayed. The configuration file allows the user to define and color code the display of sequence comparison results within Artemis. This script incorporates into the Artemis record all of the pertinent statistical and score values associated with a BLAST or sequence comparison. As the user can view the sequence comparison statistics within Artemis (*i.e.*, the identity of the target sequence hit by the query, the length and percentage identity of the hit, the E value or score), decisions are easily made as to the best results on which to base the final putative gene model.

At this point all of the collected data is available to the user to assign final putative gene models within the sequence. Artemis can be used to generate the final annotation files for submission to databases such as NCBI. Furthermore, the BLAST evidence with accompanying statistics, associated with an annotated sequence can easily be exported from Artemis to Sequin (Benson *et al*., 1998), or other software utilized to prepare GenBank entries. This allows whoever might view the data from NCBI at a later date to determine for themselves whether to trust a proposed gene model based on the given sequence identity results.

Example one: Annotating Phytochrome C

Annotation of *PHYC*[9] within a rice genomic sequence provides an example of how this annotation system allows the user to organize and display data effectively so that an accurate

---

[9] *PHY* = phytochrome gene, PHY = phytochrome protein. A specific *PHY* or PHY is designated by the letter(s) of that *PHY*/PHY or *PHY*/PHY subfamily.

annotation can be made. *PHYC* has been sequenced from a number of plants, including the model monocot rice (Basu *et al*., 2000). In this annotation example, Bacterial Artificial Chromosome (BAC) sequence OSJNBa0032E21[10] deposited in GenBank (accession number AF377947[11]), is known to contain *PHYC* through hybridization studies to a phytochrome C cDNA clone (data not shown). Evidence, from gene predictions to sequence comparisons, was developed for AF377947 as described in the preceding sections. A screen capture of the resulting visualization in Artemis is depicted in Figure 3.6. *PHYC* is on the reverse strand of AF377947, resulting in its depiction from right to left across the screen, rather than left to right.

The five gene prediction programs described in Table 3.1 were run on sequence AF377947. The results were parsed and inserted into the database as described. After reformatting the prediction outputs and loading them into Artemis for visualization it was determined that the two versions of GeneMark predicted the exact same gene model. Thus, these results are only depicted once (Figure 3.6, track 2). Nonetheless, results from the two versions of GeneMark differed from those predicted by the other three programs (Figure 3.6, tracks 1, 3, and 4). Because *PHYC* is so well-characterized and has been sequenced multiple times from a variety of plant species, only homology results with a high percentage of sequence identity and low error value are displayed in this visualization (Figure 3.6, tracks 5-7). For example, track 7 depicts those hits with sequence similarity percentages of 95% or greater when ORFs more than 50 amino acids in length are compared to the PIR-NREF database (Huang *et al*., 2003).

[10] BAC OSJNBa0032E32 was kindly provided by the Clemson University Genomics Institute (CUGI). CUGI also confirmed the presence of phytochrome C on the BAC through hybridization to a known Phytochrome C cDNA probe.

[11] AF377947 was sequenced, assembled, and annotated by A. Eastman. The complete and final sequence was submitted to GenBank on May 22, 2002.

After viewing the sequence data it becomes obvious that the gene model predicted by FgeneSH (Salamov and Solovyev, 2000) is the only accurate prediction from the five programs utilized (Figure 3.6, track 1). In addition, the results of comparison to the full-length rice cDNA database (Kukuchi *et al*., 2003) provide ample evidence of the 3' and 5' untranslated regions (UTRs) of this gene (Figure 3.6, track 5). The "repeat regions" depicted in track 8 were manually incorporated into the final annotation after AF377947 was analyzed by RepeatMasker (Smit, Hubley, and Green, 1996).

This example clearly demonstrates the ease with which the correct gene prediction can be determined and how an abundance of evidence can be combined to create the final and comprehensive annotation of the *PHYC* for submission to GenBank (Figure 3.6, track 8).

Example two: Annotating a "Problem Gene"

One of the major problems with gene prediction programs is that certain types of genes are extremely problematic when it comes to prediction (Mathé *et al*., 2002). These "problem genes" display a variety of characteristics. One type of problematically predicted gene missed by many gene prediction programs is a long gene spaced over a large (>10kb) region of the genome with multiple small exons spaced a great length apart (Mathé *et al*., 2002). An example of a gene with these characteristics is described in Figure 3.7. This gene, unlike *PHYC* described above, is located on the forward strand of AF377947 and therefore is visualized left to right on the Artemis screen capture shown in Figure 3.7. The final annotated gene (Figure 3.7, track 1) does not correspond exactly to any of the five gene prediction programs used here (Figure 3.7, tracks 4-7).

Again, all five prediction programs and sequence comparisons were performed on AF377947 and the accumulated evidence was displayed through Artemis. Results from the two

versions of GeneMark are displayed together because, again, the resulting predictions were identical. An identical result for these two versions was not observed for every prediction across the entire BAC sequence and, in areas where the predictions for the two versions of GeneMark varied, the individual predictions of both versions were evaluated.

Sequence identity evidence was not revealed during the initial comparisons of all sequences generated from the entire AF377947 data set, even though all positive results from the comparisons that could be considered non-random were included in the analysis (*i.e.*, all comparisons that yielded hits to known genes with Expect values less than 0.001). Table 3.3 describes the types of query sequences compared to various public data repositories and the explanation for the lack of resulting positive hits for this region of AF377947 (the identity of tracks 2 and 3 will be described in the next paragraph). How was the final annotation of this gene built from the varying predictions and lack of sequence evidence?

The graphic visualization of the overlapping, although not identical, gene prediction programs led to a more specific investigation into the potential for a gene being present in this region of AF377947. First, it was observed that there were multiple short exons predicted by two of the gene prediction programs, FgeneSH and GeneMark (Figure 3.7 tracks 7 and 6, respectively). Second, it was also recognized that this type of gene, one with multiple short exons spaced over a large genomic distance, is typically problematic.

The amino acid sequences of each exon from these predictions were, therefore, manually compared to the conserved domain database (CDD) of NCBI (Wheeler *et al*., 2004) and to Pfam (Bateman *et al*., 2004). The results were individually evaluated and entered into Artemis to compare with the overlapping gene predictions (Figure 3.6, tracks 2 and 3). It was revealed that certain exons from the gene predictions could be combined to yield a 100% sequence identity

match to both a MADS-box and Kbox conserved protein domain. More intensive investigation of the domains and corresponding literature led to the discovery that these two domains are commonly found together in plant transcription factors that play an important role in floral development (Reichmann and Meyerowitz, 1997; Martínez-Castilla and Alvarez-Buylla, 2003). The combination of all of this information led to the manual compilation of the final gene model (Figure 3.7, track 1). While this final annotation is simply a putative gene model, by being able to visualize the accumulated evidence, or lack thereof, and problematic nature of this type of gene, it was easily recognized that a more specific and custom investigation could be beneficial in this area of the genome. Both automated annotation pipelines and manual sorting of textual data would have missed this potential gene. In contrast, the method described here led to the fusion of different gene predictions and the identification of a putative transcription factor in this area of rice genomic sequence AF377947.

Conclusions

While the approach described here is most like Genescript and GeneMachine, it offers advantages that these other programs do not. For example, the use of a relational database to store the gene prediction and sequence homology results imparts flexibility and data storage, retrieval, and query components not offered by Genescript and GeneMachine. Like Genescript, Artemis was the annotation viewer selected for this annotation system. However, as the development of the Genescript method and subsequent publication in 2003 was performed either concurrently, or after, the development of the method described in this manuscript, the work described here was neither a duplication of nor an adaptation of the Genescript method.

The method described here was designed specifically for a small- to medium-throughput sequencing project that requires detailed and customized annotation. For any genomic

sequencing project, there are two obvious, yet diametrically opposed, annotation options; either manually curate and organize the large amounts of textual data, or employ an automated pipeline designed for whole genome projects and automated annotation. The method described here is an attempt to provide a more appropriate system for a user seeking to investigate a specific topic that pertains to a specific portion of a genome. This goal was approached in four main ways.

First, *ab initio* prediction programs and homology searches were combined to increase the accuracy of gene finding within a sequence. Second, these types of evidence are displayed in an interactive graphics-based viewer that allows the user to curate and organize all available data to facilitate annotations. Third, the method was designed with flexibility in mind so that the user can add more customized gene prediction programs or sequence homology searches. Finally, all of this information is stored in a database so that the data is easily queried and retrieved. The use of two database schemas further facilitates storage of the prediction data in a format that can be easily manipulated to customize it for other analyses. Also, if the sequence being analyzed grows in length by addition of flanking sequence, numbering of the predicted genes might change within the sequence although the predictions remain the same. Provided that the sequence has not changed in the area being analyzed, Artemis displays can be easily and quickly recalculated for the new coordinates by a script updating the database, without having to re-perform multiple analyses. The database underlying this method was designed to store information from multiple prediction programs and sequence comparisons. It was intended to be flexible enough to add easily additional programs or analyses, yet still retain a level of simplicity so that a non-computer programmer can understand and implement its design.

The perl scripts described here are appended to this thesis for those wishing to implement or adapt them for their own use.

Figures and Tables

Table 3.1: Currently Supported Prediction Programs. The Program ID is a unique number that identifies each program in the gene prediction and analysis database schema (Figure 3.1A). The Program ID, which is unique in the database, allows the addition of multiple versions of the same program to the analysis if desired. For example, GeneMark version 2.2 and the HMM version 2.2a are both supported.

Program Name	Version	Program ID	Reference
Genscan	1.0	1	Burge and Karlin, 1997
RiceHMM	2.0	2	Sakata *et al*., 1999
GeneMark	2.2	3	Borodovsky and McInich, 1993
FgeneSH	1.1	4	Salamov and Solovyev, 2000
GeneMark_HMM	2.2a	5	Lukashin and Borodovsky, 1998

Table 3.2: Generation of Artemis files. Examples of multi-fastA sequences generated in Artemis and the file names necessary to satisfy the informational needs of script header.pl, which renames headers.

File Type Description	Sequence Type	Artemis Command	File Naming Convention*
Open reading frames greater than some length (x)	Nucleotide	(1) Create/mark open reading frames/set minimum value x. (2) Write bases of selection/fasta format	SequenceName_acefilenumber_*ORF*_minimum ORFlength_*nt*_YYMMDD (Ex: seqA_1_ORF_100_nt_041001)
Open reading frames greater than some length (x)	Amino Acid	(1) Create/mark open reading frames/set minimum value x. (2) Write amino acids of selected features	SequenceName_acefilenumber_*ORF*_minimum ORFlength_*aa*_YYMMDD (Ex: seqA_1_ORF_100_aa_041001)
Exons from a gene prediction program	Nucleotide	Selected all gene predictions from each predictor (1) Write bases of selection/fasta format	SequenceName_acefilenumber_program abbreviation**_ *nt*_YYMMDD (Ex: seqA_1_GS_100_nt_041001)
Exons from a gene prediction program	Amino Acid	Selected all gene predictions from each predictor (1) Write amino acids of selected features	SequenceName_acefilenumber_program abbreviation**_ *aa*_YYMMDD (Ex: seqA_1_GS_100_aa_041001)
Whole genome sequence Nucleotide	Nucleotide	(1) Select/all bases (2) Write/bases of selection	SequenceName_acefilenumber_*SEQ*_ *nt*_YYMMDD (Ex. seqA_1_SEQ_nt_041001)
Overlapping sequences	Nucleotide	Generated with the *BREAKUP* command in RCR	SequenceName_acefilenumber_*OVL*_ *nt*_YYMMDD. (Ex. seqA_1_OVL_nt_041001)

*Those fields in italics are generic to every file of this type. Those in non-italicized font change with each file generated.

**The programs are abbreviated as follows: GeneMark v. 2.0 = GM, Genscan = GS, FgeneSH = FG, RiceHMM = RH, GeneMark v 2.2a = GH.

Table 3.3: Explanation of Lack of Sequence Comparison Results in Annotation Example Two.

Sequence Type	Database Compared To	Reason for lack of returns
nt* sequence of all ORFs greater than 50 aa** in length	EST	The ORFs in example 2 were all less than 25 amino acids in length and, therefore, were not included in the analysis.
aa sequence of all ORFs greater than 50 aa in length	PIR_NREF	Same as above.
nt sequence of exons from each gene predication program	EST	The exons were so short that any hits to them, including those that were ultimately determined to be accurate, had statistical scores below the cutoff threshold.
aa sequence of exons from each gene predication program	PIR_NREF	Same as above
nt sequence of AF377947 broken up into 3000 NT overlapping pieces	EST, KOME	The short (less than 75 nt) nature of the exons present in this gene led only to hits with a probability of being significant that was too low to be included when sequences of this size were used for comparison.
nt sequence of AF377947 in its entirety	EST, KOME	Same as above, but even more so.

* nt = nucleotide
**aa = amino acid

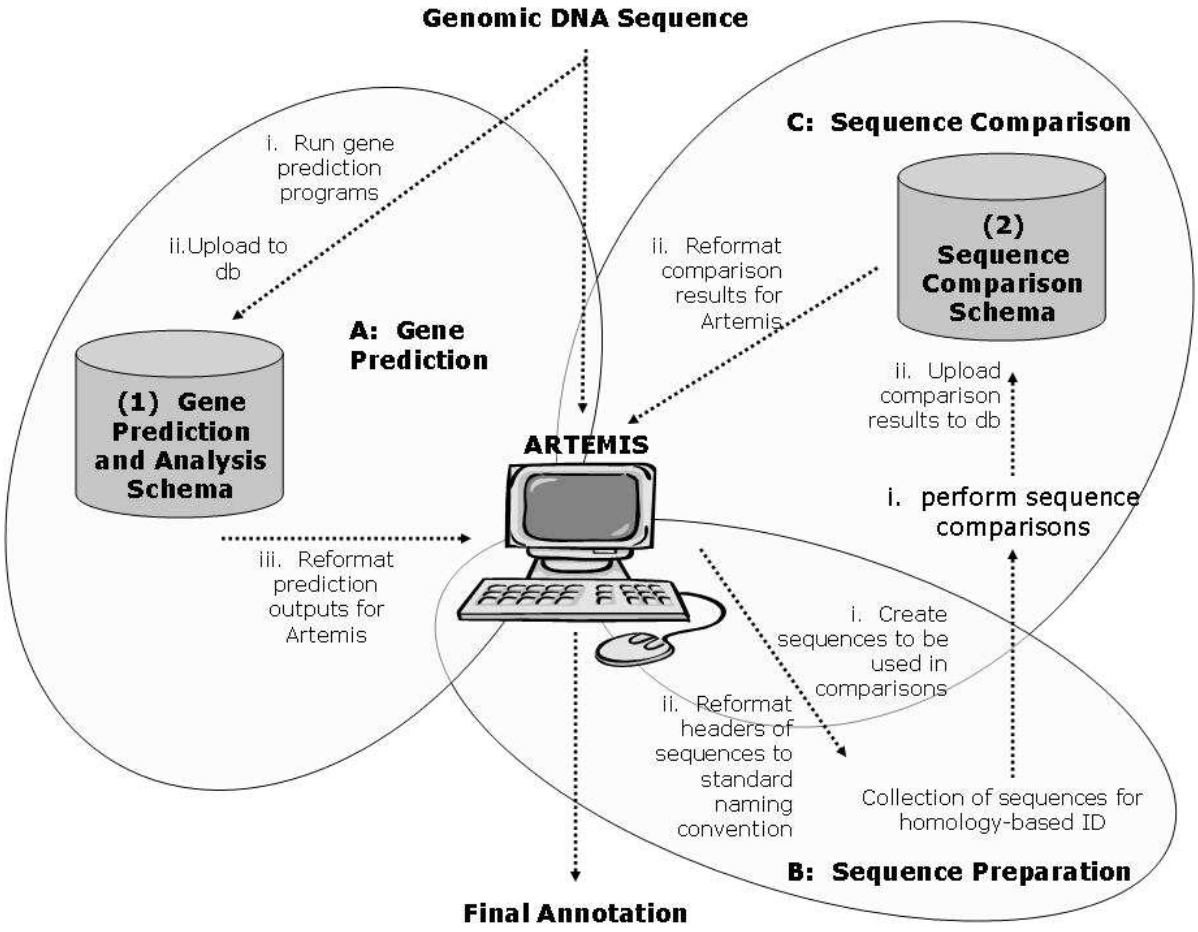


Figure 3.1: Overview of Annotation Method. (A) Gene Prediction: Predictions from multiple programs are stored in a relational database management system (RDBMS). The prediction start/stop locations are queried from the database and reformatted to EMBL format to be viewed by Artemis. (B) Sequence Preparation: Artemis functions can then be used to create any set of sequences (amino acid or nucleotide) of features of interest that require further annotation. These sequences are renamed to a standard naming convention. (C) Sequence Comparison: the sequences are then compared to databases of known gene and protein sequences. The comparison outputs are again stored in the RDBMS where they can be queried and reformatted to view in Artemis. Artemis is then used as an electronic notebook through which all lines of evidence can be viewed and annotations can be assigned to a genome sequence.

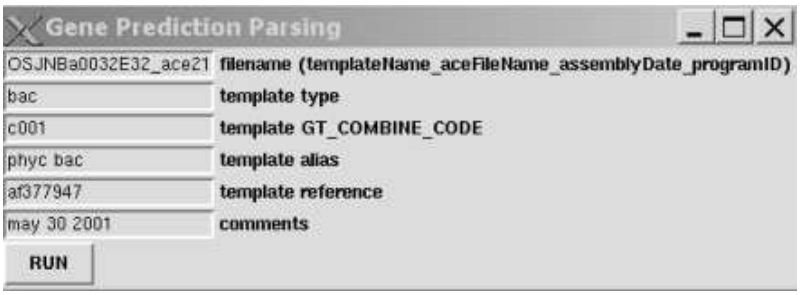
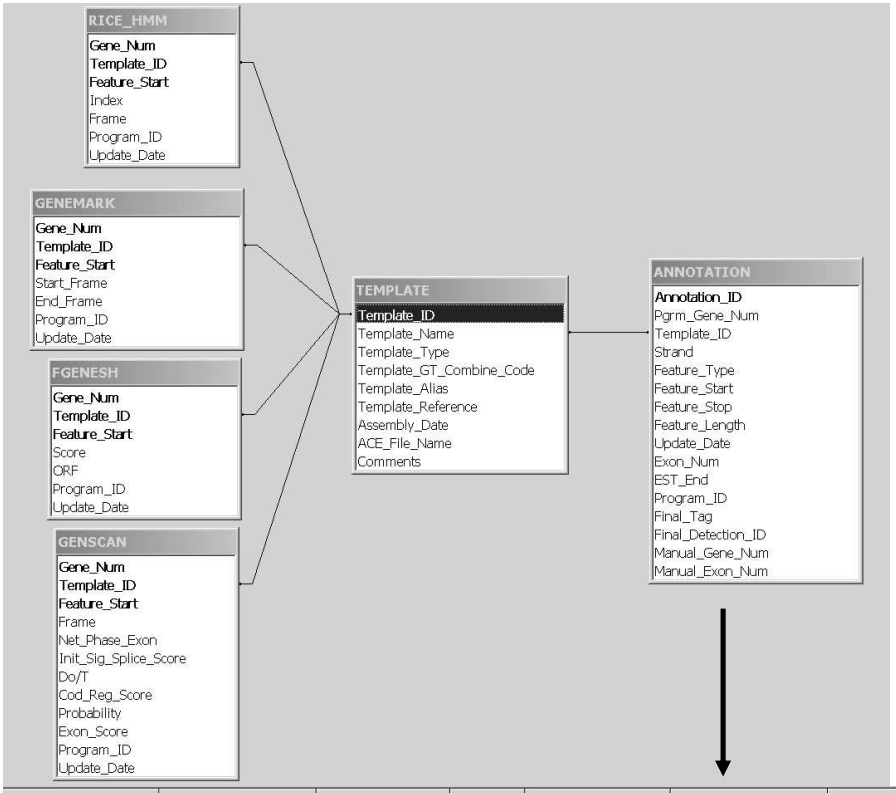


Figure 3.2: Artparse.pl Interface. A perl CGI interface activated by artparse.pl populates the gene prediction and analysis schema (see Figure 1A) with the data requested in the interface prompt. The gene prediction output file (*filename*) and information regarding the sequence being analyzed (*template type* through *comments*) is uploaded to the database through this interface. The *filename* standard naming convention contains references to both the individual sequence file of the genome sequence being analyzed and the gene prediction program ID described in Table 1. *aceFileName* refers to the actual assembly file generated by Consed from a multi-pass genomic sequencing project. Tracking the assembly file allows the genomic sequence being analyzed to be updated as it improves in length and/or quality and new prediction analyses to be easily added. *GT_COMBINE_CODE* refers to the genetic origin of the genome sequence being analyzed.



ANNOTATION_ID	PGRM_GENE_NUM	TEMPLATE_ID	STRAND	FEATURE_TYPE	FEATURE_START	FEATURE_STOP	FEATURE_LENGTH
1	1	1	plu	Terminal	946	989	44
2	2	1	plu	Initial	1404	1433	30
3	2	1	plu	Terminal	1525	1881	357
4	3	1	min	Terminal	2373	2618	246
5	3	1	min	Initial	2775	2975	201
6	4	1	plu	Initial	4243	4579	337
7	4	1	plu	Terminal	4696	5039	344
8	5	1	plu	Initial	8331	8352	22
9	5	1	plu	Internal	9175	9288	114
10	5	1	plu	Internal	9487	9771	285
11	5	1	plu	Internal	9899	9953	55
12	5	1	plu	Internal	10046	10460	415
13	5	1	plu	Initial	10882	11001	120

Figure 3.3: Gene Prediction and Analysis Schema. This schema consists of one table per prediction program, allowing additional programs to be easily incorporated into the analysis, a "template" table to track the nature of the nucleotide sequence being analyzed, and a composite "annotation" table that stores the start/stop location of all predicted exons. One "GeneMark" table serves the storage needs of both versions of GeneMark (see Table 1) incorporated into this method. The annotation table can also support the storage of all final annotated features within a sequence (see table enlarged below under the arrow); although an automatic method of populating this table after the final annotation does not currently exist.

Figure 3.4: Gene Prediction Comparisons in Artemis. This screen capture illustrates how Artemis displays the results from different gene prediction programs. Regions where gene predictions overlap or significantly agree are easily identified for targeted annotation (see circled regions).

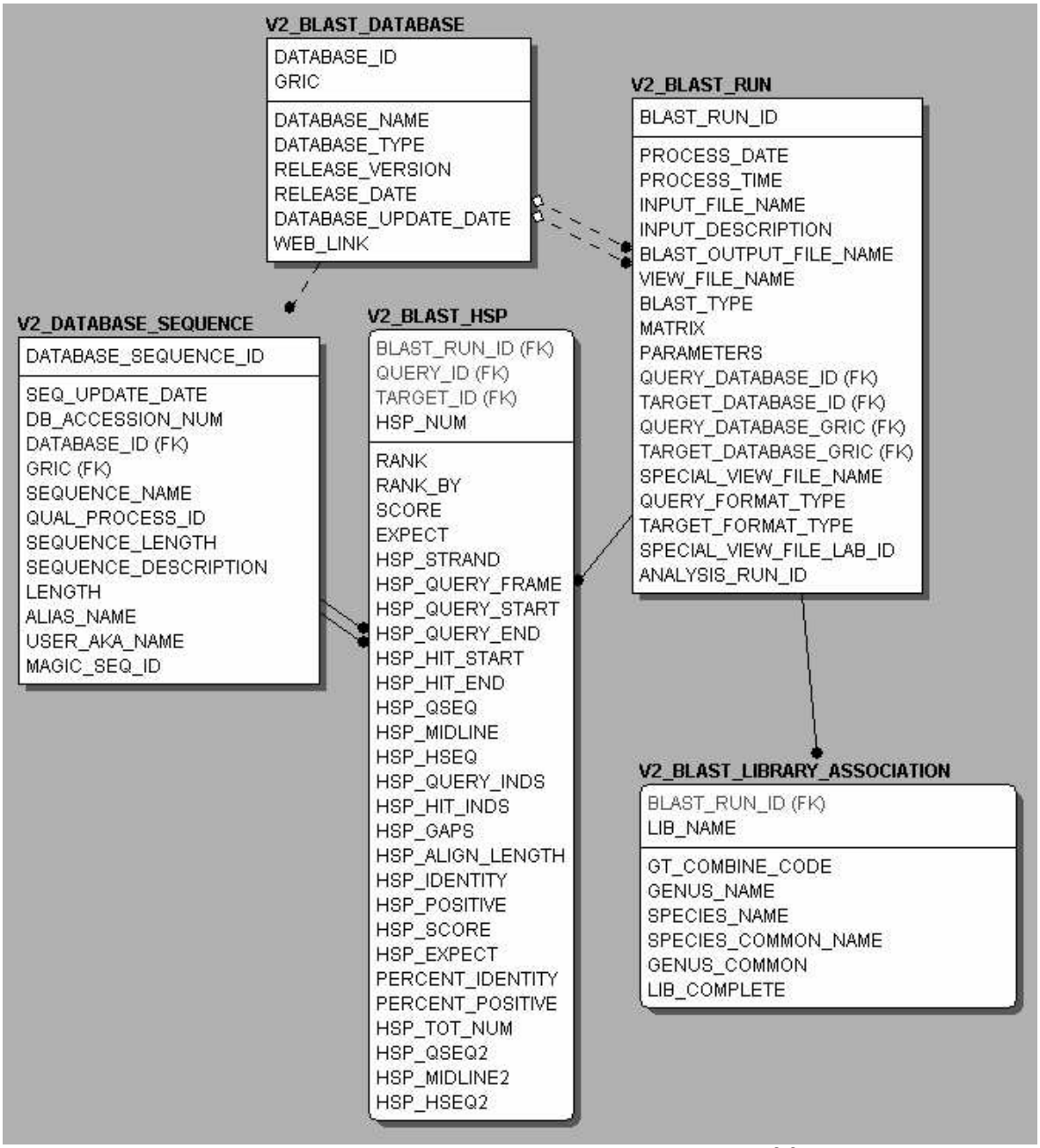


Figure 3.5: Sequence Comparison Schema. This schema consists of five tables which hold the following: information regarding the databases to which sequences were compared (V2_BLAST_DATABASE), the particulars and parameters of each comparison (V2_BLAST_RUN), the results of the comparisons (V2_BLAST_HSP), information about each query sequence used for the comparison (V2_DATABASE_SEQUENCE), and information about the types of query sequences used in the comparisons (V2_BLAST_LIBRARY_ASSOCIATION)[12].

[12] This schema was designed as described in Cordonnier-Pratt *et al*., 2004.

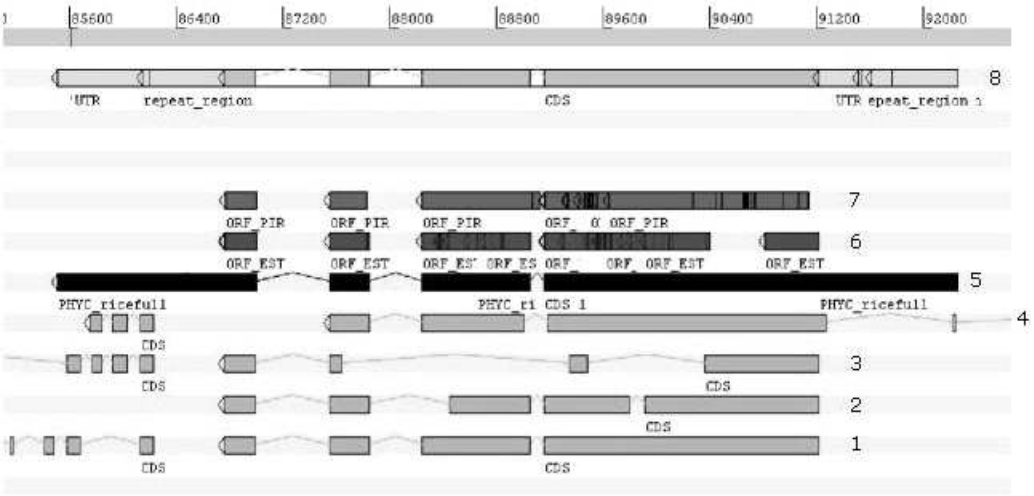


Figure 3.6: Phytochrome C in AF377947. The numbering along the top of the image refers to the position of this gene within the BAC sequence in nucleotides (~ 85600-92000 on the reverse strand of AF377947). Tracks 1-4 denote the gene prediction programs used (1 = FgeneSH, 2 = both versions of GeneMark, 3 = Genscan, and 4 = RiceHMM). Tracks 5-7 denote the sequence comparison evidence shown here (5 = FgeneSH exons plus 2000 nt upstream and downstream of the initial and terminal exons compared to the KOME rice full length cDNA database, 6 = nucleotide sequences of all open reading frames greater than 50 amino acids compared to the EST database from NCBI, and 7 = amino acid sequences of all open reading frames greater than 50 amino acids compared to the PIR-NREF database. Track 8 denotes the final Phytochrome C gene model annotated and submitted to GenBank.

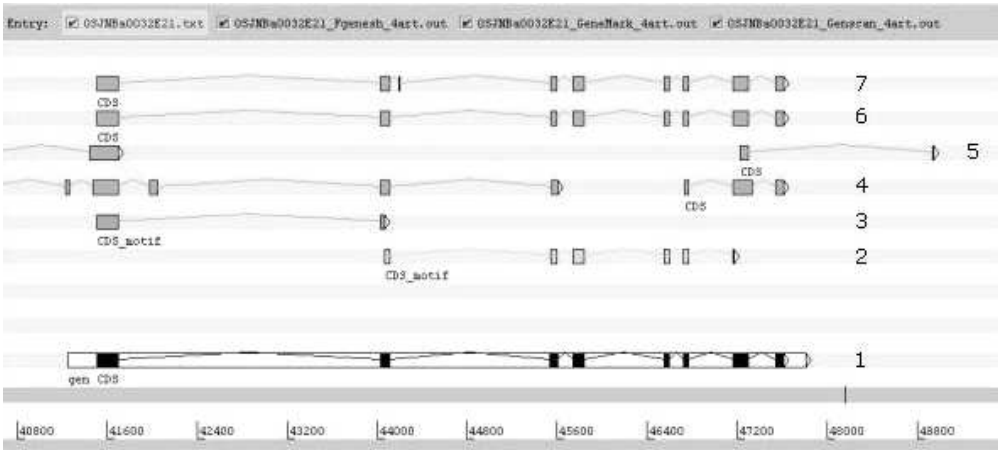


Figure 3.7: A "Problem Gene" in AF377947. The putative gene model described here (track 1) was annotated in the absence of gene prediction agreement or sequence homology evidence generated through normal automated means. None of the gene prediction programs agree (track 7 = FgeneSH, 6 = GeneMark, 5 = Genscan, and 4 = RiceHMM). Tracks 2 and 3 are the motifs discovered through individual manual comparison to conserved protein domain databases.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**: 403-410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., *et al*,. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**: 3389-3402.

Basu, D., Dehesh, K., Schneider-Poetsch, H.J., Harrington, S.E., McCouch, S.R., Quail, P.H. (2000) Rice PHYC gene: structure, expression, map position, and evolution. *Plant Mol. Biol.*, **44**: 27-42.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moon, S., Sonnhammer, E.L., Studholme, D.J., Yeats, C., Eddy, S.R. (2004) The Pfam protein families database. *Nucleic Acids Res*., **Database issue**, D 138-41.

Benson, D.A., Boguski, M.S., Lipman, D.J., Ostell, J., Ouellette, B.F. (1998) GenBank. *Nucleic Acids Res*. **26**: 1-7.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. (2003) GenBank. *Nucleic Acids Res.*, **31**, 23–27.

Birney, E., Clamp, M., and Durbin, R. (2004) Genewise and genomewise. *Genome Res.*, **14**: 988-995.

Borodovsky, M. and McIninch, J. (1993) GeneMark: parallel gene recognition for both DNA strands. *Computers & Chemistry*, **19**: 123-133.

Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol*. **268**: 78-94.

Clamp, M., Andrews, D., Barker, D., Bevan, P., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Hubbard, T., Kasprzyk, A., Keefe, D., Lehvaslaiho, H., Iyer, V., Melsopp, C., Mongin, E., Pettett, R., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik. I,, Birney. E. (2003) Ensembl 2002. *Nucleic Acids Res*., **31**: 98-42.

Cordonnier-Pratt, M.-M., Liang, C., Wang, H., Kolychev, D.S., Sun, F., Freeman, R., Sullivan, R., Pratt, L.H. (2004) MAGIC Database and interfaces: an integrated package for gene discovery and expression. *Comparative and Functional Genomics*, **5**: 268-275.

Dowell, R., Jokerst, R., Day, A., Eddy, S.R., Stein, L. (2001) The distributed annotation system. *BMC Bioinformatics*. **2**: 7.

Florea, L., Hartzell, G. Zhang, Z., Rubin, G.M., Miller, W. (1998) A computer program for aligning cDNA sequence with a genomic DNA sequence. *Genome Res*., **8**: 967-974.

Gordon, D., Abjian, C., Green, P. (1998) Consed: a graphical tool for sequence finishing. *Genome Res*., **8**; 195-202.

Huang, H., Barker, W.C., Chen, Y., Wu, C.H. (2003) iProClass: an integrated database of protein family classification, function, and structure information. *Nucleic Acids Res*., **31**: 390-392.

Hudek, A., Cheung, J., Boright, A.P., and Scherer, S.W. (2003) Genescript: DNA sequence annotation pipeline. *Bioinformatics* **19**: 1177-1178.

Iseli, C., Jongeneel, C.V., Bucher, P. (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc Int Conf Intell Syst Mol Biol*.,138-148.

Kikuchi, S., Satoh, K., Nagata, T., Kawagashira, N., Doi, K., Kishimoto, N., Yazaki, J., Ishikawa, M., Yamada, H., Ooka, H., Hotta, I., Kojima, K., Namiki, T., Ohneda, E., Yahagi, W., Suzuki, K., Li, CJ., Ohtsuki, K., Shishiki, T., Otomo, Y., Murakami, K., Iida, Y., Sugano, S., Fujimura, T., Suzuk,i Y., Tsunoda, Y., Kurosaki, T., Kodama, T., Masuda, H., Kobayashi, M., Xie, Q., Lu, M., Narikawa, R., Sugiyama, A., Mizun, K., Yokomizo, S., Niikura, J., Ikeda, R., Ishibiki, J., Kawamata, M., Yoshimura, A., Miura, J., Kusumegi, T., Oka, M., Ryu, R., Ueda, M., Matsubara, K., Kawai, J., Carninci, P., Adachi, J., Aizawa, K., Arakawa, T., Fukuda, S., Hara, A., Hashizume, W., Hayatsu, N., Imotani, K., Ishii, Y., Itoh, M., Kagawa, I., Kondo, S., Konn,o H., Miyazaki, A., Osato, N., Ota, Y., Saito, R., Sasaki, D., Sato, K., Shibata, K., Shinagawa, A., Shiraki, T., Yoshino, M., Hayashizaki, Y., Yasunishi, A.; Rice Full-Length cDNA Consortium; National Institute of Agrobiological Sciences Rice Full-Length cDNA Project Team; Foundation of Advancement of International Science Genome Sequencing & Analysis Group; RIKEN. (2003) Collection, Mapping, and Annotation of over 28,000 cDNA clones from japonic Rice. *Science*, **301**: 376-379.

Krogh, A. (1997) Two methods for improving performance of an HMM and their application for gene finding. *Proc. Intl. Conf. Intell. Syst. Mol. Biol.* **5**: 179-186.

Lewis, S.E., Searle, S.M., Harris, N., Gibson, M., Lyer, V., Richter,J., Wiel, C., Bayraktaroglir, L., Birney, E., Crosby, M.A., Kaminker, J.S., Matthews, B.B., Prochnik, S.E., Smithy, C.D., Tupy, J.L., Rubin, G.M., Misra, S., Mungall, C.J., Clamp, M,E. (2002) Apollo: a sequence annotation editor. Genome Biol., **3**: RESEARCH0082.

Lukashin, A.V. and Borodovsky, M. (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res*., **26**: 1107-1115.

Makalowska, I., Ryan, J.F., Baxevanis, A.D (2001) GeneMachine: a gene prediction and sequence annotation. *Bioinformatics* **17**: 843-844.

Martínez-Castilla, L. P. and Alvarez-Buylla, E. R.(2003) Adaptive evolution in the *Arabidopsis* MADS-box gene family inferred from its complete resolved phylogeny. *Proc Natl Acad Sci U S A.* **100**: 13407–13412.

Mathé, C., Sagot, M.-F., Schiex, T., and Rouzé, P. (2002) Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res*., **30**: 4103-4117.

Mural, R. J., Einstein, J.R., Guan, X., Mann, R.C., and Uberbacher, E.C. (1992) An artificial intelligence approach to DNA sequence feature recognition. *Trends Biotechnol*., **10**: 67-69.

Pavy, N., Rombauts, S., Dehais, P., Mathe, C., Ramana, D.V., Leroy, P., Rouze, P**.** (1999) Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences. *Bioinformatics*. **15**: 887-899.

Riechmann, J. L. and Meyerowitz, E. M. (1997) MADS domain proteins in plant development. *Biological Chemistry*. **378**: 1079-1101.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., Barrell, B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics* **10**: 944-945.

Sakata, K., Nagamura, Y., Numa, H., Antonio, B.A., Nagasaki, H., Idonuma, A., Watanabe, W., Shimizu, Y., Horiuchi, I., Matsumoto, T., Sasaki, T., Higo, K. (2002) RiceGAAS: an automated system for rice genome annotation. *Nucleic Acids Res*. **30**: 98-102.

Sakata, K. Nagasaki, H., Idonuma, A., Waki, K., Kise, M., Sasaki, T (1999) A computer program for prediction of gene domain on rice genome sequence. *The 2nd Georgia Tech International Conference on Bioinformatics*, **Abstracts** p.78.

Salamov, A. and Solovyev, V. (2000) *Ab initio* gene finding in Drosophila genomic DNA. *Genome Res*., **10**: 516-522.

Schoof, H. and Karlowski, W. (2003) Comparison of rice and *Arabidopsis* annotation. *Curr. Opin. Plant Biol.* **6**: 106-112.

Smit, AFA, Hubley, R & Green, P. (1996-2004) *RepeatMasker Open-3.0*. <http://www.repeatmasker.org>.

Solovyev, V.V., Salamov, A.A., Lawrence, C.B. (1995) Identification of human gene structure using linear discriminant functions and dynamic programming. *Ismb*. **3**: 367-375.

Ware, D.H., Jaiswal, P., Ni, J., Yap, I.V., Pan, X., Clark, K.Y., Teytelman, L., Schmidt, S.C., Zhao, W., Chang, K., Cartinhour, S., Stein, L.D., McCouch, S.R. (2002) Gramene, a tool for grass genomics. *Plant Physiol*. **130**: 1606-1613.

Wheeler, D.L., Church, D.M., Edgar, R., Federhen, S., Helmberg, W., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Sequiera, E., Suzek, T.O., Tatusova, T.A., Wagner, L. (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res*., **Database issue,** D35-40.

Womble, D.D. (2000) GCG: The Wisconsin Package of sequence analysis programs. *Methods Mol. Biol*., **132**: 3-22.

Zhang, M.Q. (1997) Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc. Natl. Acad. Sci., USA*, **94**: 565-568.

CHAPTER 4

THE PHYTOCHROME GENE FAMILY AND FLANKING DNA IN RICE

Introduction

Plants are immobile organisms that must be able to detect and respond to changes in their environment in order to thrive. To this end, plants have evolved complex sensory mechanisms allowing them to respond to their surroundings, specifically in regards to the detection of ambient light. In addition to utilizing light quality and quantity to initiate response mechanisms, plants employ solar radiation as an energy source to drive photosynthesis. A number of photosensors are involved in the photosensory detection of light. They are linked directly to many developmental stages of a plant's life cycle. Probably the best characterized photoreceptors are the phytochromes, which absorb light preferentially in the red (660 nm) and far-red (730 nm) regions of the visible spectrum (Borthwick *et al*., 1948).

Given the importance of light detection among members of the plant kingdom, both the function and evolutionary history of the phytochromes have been extensively studied in plants such as Arabidopsis, the model dicot and the first plant for which an entire genome sequence was published (Sharrock and Quail, 1989; Dehesh *et al*., 1993; Nagatani, Reed, and Chory, 1993; Parks and Quail, 1993; Whitelam *et al*., 1993; Reed *et al*., 1993; Clack *et al*., 1994; Aukerman *et al*., 1997; Davis *et al*., 1999; Devlin *et al*., 1999; Muramoto *et al*., 1999; The Arabidopsis Genome Initiative, 2000). Phytochromes in all higher plants appear to have the same basic structure, a homodimer comprised of two 124-kDa subunits (Pratt, 1982; Lagarias, 1985). Characterization of the phytochrome gene families in both angiosperm (Sharrock and Quail, 1989; Howe *et al*., 1989; Clack *et al*., 1994; Pratt *et al*., 1997; Pasentsis *et al*., 1998) and non-angiosperm (Quail, 1991; Kolukisaoglu *et al*., 1993; Mathews, Lavin, and Sharrock, 1995; Mathews and Sharrock, 1996) plants has demonstrated that the gene family has become

substantially larger and more complex during evolution from its cyanobacterial progenitor (Kolukisaoglu *et al*., 1993; Mathews and Sharrock, 1996).

The most common hypothesis of *PHY*[1] family expansion submits that the presumed progenitor *PHY* duplicated to yield the *PHYA/C/F* and *PHYB/D/E* subfamilies prior to the divergence of conifers from angiosperms (Donoghue and Mathews, 1999; Mathews and Sharrock, 1996) (Figure 4.1). Before the monocot/dicot divergence in angiosperms the *PHYA/C/F* subfamily then underwent a second duplication event ~200 Mya (Wolfe *et al.*, 1989) to yield the *PHYA* and *PHYC/F* subfamilies (Donoghue and Mathews, 1999). This duplication yielded the three-membered *PHYA, PHYB,* and *PHYC* family characteristic of current-day monocots (genes denoted by * in Figure 4.1). Subsequent duplication/divergence events in dicots are thought to have given rise to *PHYC, PHYF, PHYE, PHYB,* and *PHYD* as illustrated. One basic tenet of this hypothesis is that while the *PHY* family in dicots appears to be separated into four or five subfamilies (*PHYA, B/D, C, E,* and *F*) (Clack *et al.*, 1994; Mathews and Sharrock, 1996), monocots have representatives from only *PHYA, PHYB,* and *PHYC* (Mathews and Sharrock, 1996).

Although it is an accepted phenomenon among phytochrome researchers that the phytochrome family has expanded from a cyanobacterial progenitor, a predominant question for phytochrome evolutionary scientists is *why* and *how* did this family expand? Approximately 140-190 Mya[2] the dominant plant lineages present in the environment changed upon the introduction of angiosperms to the ecosystem (Sanderson and Doyle, 2001). One hypothesis states that early angiosperms would have benefited from any evolutionary modification that

[1] *PHY* = phytochrome gene, PHY = phytochrome protein. A specific *PHY* or PHY is designated by the letter(s) of that *PHY*/PHY or *PHY*/PHY subfamily.

[2] Mya: Million years ago

made them able to thrive and establish themselves in a dimly lit environment (Mathews, Burleigh, and Donoghue, 2003). The evolution of multiple phytochromes would facilitate the detection of surrounding light levels and the wavelength distribution of that light and would therefore impart an advantage over those plants having only one or a few phytochromes.

If increasing functional diversity is the driving force behind *why* the *PHY* family expanded, the next logical question is *how* did the family expand? Did the phytochrome genes duplicate individually or did surrounding regions of the genome duplicate as well? Were there further translocations or deletions of phytochrome genes after these duplication events? Detecting the mechanisms of *PHY* family expansion that presumably resulted in plants evolving from organisms with one to many phytochromes is not self-evident due to the large amount of evolutionary time that has passed since the hypothesized duplication events. Furthermore, investigating these mechanisms from a genomic perspective is directly related to the currently available *PHY*-containing genome sequences of plants. At the inception of this study, the only *PHY*-flanking genomic sequence publicly available from a higher plant species was from the dicot Arabidopsis (The Arabidopsis Genome Initiative, 2000), although other *PHY*-containing sequences have been made available during the course of this study (Morishige *et al*., 2002).

Overview of Study

The study presented here sought to characterize the *PHY* gene family and the genomic sequence flanking these *PHY* in the monocot plant *Oryza sativa* (rice). This study can be broken down into three main components: (I) sequence the *PHY* family and flanking genomic sequences in rice, (II) extensively and accurately annotate these genomic sequences and (III) compare these

genome sequences of rice to each other and to other *PHY*-containing genomic sequences to investigate possible mechanisms of *PHY* family expansion in plants.

(I) Sequence the PHY *Family and Flanking Genomic Sequences*

Rice was chosen as the monocot from which to characterize the *PHY* family for a number of reasons. First, it is obviously important as an international food crop. The cereals, a family of which rice is a member, account directly or indirectly for at least 70% of the world's food production. Second, the relatively small genome size of 430 megabases (Arumunganathan and Earle, 1991) contains fewer repeat sequences than other plants with larger genomes and is therefore easier to sequence. Third, the commencement of this study coincided with the initial phases of the sequencing portion of the International Rice Genome Sequencing Project (http://btn.genomics.org.ch:8080/rice/). Sequencing these three regions of rice was intended to benefit the international effort as the regions sequenced here were located in areas scheduled to be addressed in the last stages of the public sequencing effort. Fourth, bacterial artificial chromosomes (BACs) known to contain each of the different *PHY* found in rice had already been identified and were available for sequencing.

Rice, like most monocots, contains three *PHY* (*PHYA*, *B*, and *C*). Each BAC sequenced during this study contained one of these *PHY*. The BACs also contained flanking genomic sequences both upstream and downstream of *PHY*. This raw genome sequence, however, yields little useful information in and of itself. The genomic sequence must be accurately and extensively annotated to characterize the *PHY* family and flanking genomic DNA.

(II) Annotate the BAC Sequences

The BAC sequences were annotated using the approach described in Chapter 3. This approach was designed specifically for a researcher involved in low to moderate throughput DNA

sequencing who desires a detailed, manually curated annotation that can be individualized to address a specific biological question.

(III) Compare these Sequences to Each Other and to Other PHY*-containing Sequences to Investigate Possible Mechanisms of* PHY *Family Expansion*

The annotated BAC sequences were compared to each other and to genomic sequences surrounding *PHY* in *Arabidopsis* to determine whether any DNA flanking *PHY* was duplicated and retained within the rice genome and between the rice and *Arabidopsis* genomes. Because these regions of the rice genome diverged at least 120 Mya, and rice and *Arabidopsis* diverged 120-200 Mya (Wolfe *et al*., 1989; Donoghue and Mathews 1999), a combinatorial approach including local and global alignment algorithms was employed to facilitate the detection of any regions of conservation among the sequences. In this study, BL2SEQ (Tatusova and Madden, 1999), BLASTN (Altschul *et al*., 1990), BLASTX (Salse *et al*., 2002), and *AVID* (Bray *et al*., 2003) were used to compare the *PHY*-containing BAC sequences to one another.

Methods

Shotgun Sequence Three PHY*-containing BACs from Rice*

Three BACs were chosen from those kindly provided by the Clemson University Genome Institute[3] based on hybridization intensity to known phytochrome cDNA probes. BAC DNA was isolated using a standard large-scale alkaline lysis method (Sambrook *et al*., 1989; Roe *et al*., 1996). The DNA was randomly sheared by nebulization at 55 kPa (8 p.s.i.) for 2 min (Roe *et al*., 1996). A 1% (w/v) low-melt agarose gel was used to size-fractionate the fragments into two pools of 1-2 kb and 2-4 kb. The fragments were purified from the gel and ligated into

[3] The Clemson University Genomics Institute (CUGI) website can be found at http://www.genome.clemson.edu/.

dephosphorylated pUC18 vector. *E. coli* was transformed with the plasmids by electroporation (BioRad, 1994) and plated on selective medium containing 100 mg/ml ampicillin. Randomly selected subclones were inoculated into culture medium in 96-well plates and grown overnight. The majority of the plasmids (~90%) were isolated using a standard double alkaline lysis procedure (Roe *et al*., 1996). The remainder were purified using a GeneMachines® Revolution Prep™ Machine, a robotic workstation for automated centrifugation-based plasmid DNA purification (Genomic Solutions, San Carlos, CA). A sufficient number of different plasmids were isolated to sequence the BACs to approximately 10-fold coverage, assuming an average BAC length of 150 kb.

Plasmids were sequenced in both directions using M13-21 (5' TGTAAAACGACGGCCAGT 3') and JenRev (5' CAGGAAACAGCTATGACC 3') primers. Sequencing was performed with an ABI Prism BigDye™ Terminator Cycle Sequencing Ready Reaction Kit (versions 2.0 and 3.0) in both 96- and 384-well format. Each sequencing reaction included 4 µl of plasmid DNA (~100-300 ng total DNA), 44 pmol of primer, 0.70 µl of BigDye™, 0.58 µl ddH_2O, 0.34 µl DMSO, 1.4 µl 5x reaction buffer (5x = 10 mM $MgCl_2$, 400 mM Tris-Cl pH 9.0), resulting in a total reaction volume of 7.2 µl. Thermal cycling reactions were performed with a 96-well GeneAmp® PCR System (Applied Biosystems, catalog no. N805-0200) and an Autolid Dual 384-well GeneAmp® PCR System 9700 (Applied Biosystems, catalog no. N805-0400). Cycling parameters were 99 cycles of 96°C, 10s; 50°C, 5s; 60°C, 4 min. Thermal cycling reaction products were purified by filtration through Sephadex G-50 (Amersham Pharmacia, catalog no. 17-0043-02) in 96-well multiscreen filter plates (Fisher, catalog no. MADVN6550), or in genCLEAN 384-well filter plates (Genetix, catalog no. K1025,

New Milton, Hampshire, UK). The purified extension products were separated on an ABI Prism® 3700 DNA Analyzer (Applied Biosystems).

Bases were called with phred version 0.990329 and sequences were assembled with phrap version 0.990329 with the minmatch and maxmatch parameters changed from default to 30 and 55, respectively. Sequence gaps were closed by one of three methods. Subclones spanning the gap were identified and sequenced with (1) universal primers or (2) custom primers designed to extend available insert sequence. Alternatively, (3) subclones spanning the gap or at the flanking ends of gaps were subject to transposon mediated sequencing with an EZ::TN <KAN-2> Transposome ™ Kit (Epicentre, catalog no. TMS99K2, Madison, WI; Shevchenko *et al*., 2002).

Final sequence assemblies were verified by comparing a restriction digest of the isolated BAC DNA to an electronically generated restriction digest profile of the assembled sequence (data not shown).

Annotation of the Three PHY*-containing Genome Sequences*

The assembled BAC sequences were annotated to identify potential genes upstream and downstream of the known *PHY* using the perl scripts, database schemas, visualization tools, and the overall bioinformatic approach described in Chapter 3. The annotation strategy described here can be divided into five main tasks: (A) use multiple algorithms to predict probable gene locations within the BAC sequences, (B) compare the BAC sequences, predicted gene sequences, and other sequences of interest from within the BAC to databases of known amino acid and nucleotide sequences, (C) use comparisons to protein domain databases to refine annotations and increase knowledge regarding gene function, and investigate the BAC sequences

for (D) transposons and retrotransposons and (E) remnants of ancient phytochrome genes or pseudogenes.

(A) The following gene prediction programs were used to identify potential coding regions in the assembled BAC sequences: RiceHMM version 2 (Sakata *et al*., 1999), FGeneSH version 1.1 (Salamov and Solovyev, 2000), Genscan version 1 (Burge and Karlin, 1997), and GeneMark versions 2.2 (Borodovsky and McIninch, 1993) and 2.2a (Lukashin and Borodovsky, 1998).

(B) Six different files containing sequences in fastA format were produced for each BAC (Table 4.1). FastA files containing the nucleotide and amino acid sequences of the gene predictions were produced by the "write/amino acids of selection" and "write/bases of selection" commands in Artemis (Table 4.1: file numbers 1 and 2). FastA files containing the nucleotide and amino acids sequences of all open reading frames (ORFs) longer than 50 amino acids were produced with Artemis using the "create/mark open reading frames" command followed by the "write/amino acids of selection" or "write/bases of selection" commands (Table 4.1: file numbers 3 and 4). Each BAC sequence was divided into overlapping 3-kb sequences (Table 4.1; file number 5) by the "breakup" command within the GCG package distributed by the Genetics Computer Group (Wisconsin Package version 10.2, Madison, Wisconsin). Finally, the BAC sequence itself was converted to fastA format (Table 4.1: file number 6) by the "tofasta" command in GCG (Wisconsin Package version 10.2, Madison, Wisconsin.).

These six fastA files were then compared to multiple databases and public sequence repositories (Table 4.2). Nucleotide sequences were compared to the KOME[4] database of full-length rice cDNA sequences and the est_others and nonredundant (NR) nucleotide sequence

databases from the National Center for Biotechnology Information (NCBI). In addition to comparison to these three databases, the nucleotide sequences of the assembled BACs were also compared, in their entirety, to the high-throughput genomic sequences (htgs) and swiss_prot_plus databases from NCBI. Amino acid sequences were compared to the PIR_NREF[5] and Pfam[6] databases as well as swiss_prot_plus from NCBI.

(C) The amino acid sequences of all regions likely to contain a gene were compared to the conserved domain database (CDD) (Marchler-Bauer *et al*., 2003 and 2005) through the web interface accompanying this database (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) for the identification of any protein domains present within the sequences.

(D) The presence/absence of transposons and retrotransposons in the assembled BAC sequences was also investigated. A collection of LTR retrotransposon sequences identified in the rice genome by the algorithm LTR_STRUC (McCarthy *et al*., 2002) and transposon sequences mined from rice sequences available in GenBank and from the Monsanto collection was kindly provided by Gene McCarthy and John McDonald[7]. A database of these sequences was prepared using the "formatdb" command in GCG. The three assembled BAC sequences were compared to this database using the BLASTN[8] algorithm. Moreover, three databases of the assembled BAC sequences divided into 3-kb overlapping sequences were prepared using "formatdb" and

---

[4] KOME: downloaded from http://cdna01.dna.affrc.go.jp/cDNA/CDNA_main_front.html
[5] PIR_NREF: downloaded from ftp.pir.georgetown.edu/pir_databases/nref.
[6] Pfam: downloaded form ftp://ftp.sanger.ac.uk/pub/databases/Pfam/.
[7] University of Georgia, Department of Genetics, Life Sciences Building, Athens, GA 30602
[8] Throughout the remainder of this chapter, unless otherwise noted, the BLASTN algorithm was run with the following default and non-stringent parameters to identify sequence similarity:
Default: reward for match = 1, penalty for mismatch = 2, cost to open a gap = 5, cost to extend the gap = 2, word size = 11

each of the transposon/retrotransposon sequences present in the McCarthy/McDonald collection was compared to these databases with BLASTN[9]. This approach of creating databases that only contain the sequences being searched for yields error values (E-values) that reflect the actual statistical significance of an LTR, transposon, or retrotransposon being present in the assembled BAC sequences. All results with E-values less than 0.01 were manually evaluated against other sequence similarity evidence using Artemis.

(E) This same approach was used to search for low level homology to known phytochrome genes that might be remnants of ancient *PHY* still present in the genomic sequence surrounding current functional phytochrome genes. The NR database at NCBI was queried for the complete coding sequences of any plant phytochrome genes. Full-length phytochrome sequences from rice, sorghum, and Arabidopsis were then manually compiled from the results of this query using the "fetch" command in GCG (Table 4.3). As described above, a database containing only these sequences was then created to which the BAC sequences were compared. The individual *PHY* sequences were also compared to the database compiled from the BACs divided into 3-kb overlapping sequences.

Comparing PHY*-containing Genomic Sequences*

The three assembled *PHY*-containing BAC sequences were compared to each other in order to investigate any detectable sequence identity between them that might shed light on the mechanisms of *PHY* family expansion in rice. In addition to comparing *PHY*-containing genome

[9] Reward for match = 1, penalty for mismatch = 2, cost to open a gap = 3, cost to extend the gap = 1, and word size = 8

sequences within the rice genome, the genomic rice sequences produced here were compared to other *PHY*-containing genome sequences from other plants.

These sequences, which are enumerated in Table 4.3, were used to query the htgs database at NCBI for assembled or draft BAC/PAC genome sequences that showed sequence similarity to any of these genes using BLASTN. This same set of phytochrome gene sequences was used to query MaizeDB (Laurence *et al*., 2004) and MatDB (Schoof *et al*., 2004) for genome sequences specific to these organisms that contain *PHY* in maize and Arabidopsis, respectively. In addition to the three rice BAC sequences produced here, six other genomic sequences containing a *PHY* (Table 4.4) were identified by this approach. These 9 sequences yield 21 pairs for comparative analyses (Table 4.5).

A combinatorial approach including local and global alignment algorithms was employed to facilitate detection of regions of conservation among the sequences. Each of the 21 sequence pairs was compared in the following manner.

(1) One sequence from each pair was reformatted into its own database using the "formatdb" command in the GCG package (Wisconsin Package version 10.2, Madison, Wisc.). The other sequence was compared to this database using BLASTN. This analysis was repeated using the alternative sequence in the pair as the reformatted database.

(2) The sequences were compared to each other in pair-wise fashion using BL2SEQ[10].

(3) The sequences were compared in pair-wise fashion using *AVID* through an online interface (Bray *et al*., 2003, with supplementary interface located at http://baboon.math.berkeley.edu/mavid/).

[10] BL2SEQ parameters: reward for match = 1, penalty for mismatch = 2, cost to open a gap = 5, cost to extend the gap = 2, word size = 11

(4) The amino acid sequences of all annotated ORFs from rice *PHY*-containing BACs OSJNBa0031O09, OSJNBa0016B07, and OSJNBa0032E21 were made into databases again using "formatdb". The remaining genomic BAC sequences were divided into 3-kb overlapping nucleotide sequences using "breakup" and compared to the databases using BLASTX[11] (Salse *et al*., 2002).

(5) Entire BAC sequences were compared to the databases of all annotated ORFs from each *PHY*-containing BAC using BLASTX without being divided into 3-kb overlapping sequences.

The results of the BLASTN and BL2SEQ analyses were reformatted by MSPCrunch (Sonnhammer and Durbin, 1994) to be viewed with the Artemis Comparison Tool (ACT; http://www.sanger.ac.uk/Software/ACT/). *AVID* results were initially viewed with *VISTA* (Bray *et al*., 2003), an accompanying program designed specifically for this purpose. Both *AVID* and BLASTX results were subsequently reformatted for visualization with ACT.

Results

Shotgun Sequencing Three PHY*-containing BACs from Rice*

Twelve, twenty-one, and eleven BACs putatively containing *PHYA*, *B*, and *C*, respectively, were kindly provided by the Clemson University Genomics Institute. The BACs were determined to contain a *PHY* based on the ability of a phytochrome probe to hybridize to them. One BAC per *PHY* was then selected for complete sequencing. The three *PHY*-containing BACs selected for this study were OSJNBa0031O09 (*PHYA*), OSJNBa0016B07 (*PHYB*), and OSJNBa0032E21

[11] cost to open a gap = 11, cost to extend the gap = 1, word size = 3, matrix = blossum62

(*PHYC*). Results have been deposited in GenBank on June 04, 2002 (*PHYA*[12]), March 09, 2003 (*PHYB*[13]), and May 22, 2002 (*PHYC*[14]).

PHYA, *B*, and *C* are located 95.3, 22.2, and 113.3 cM, respectively, from the top of rice chromosome 3 based on the Cornell RFLP genetic map (Causse *et al*., 1994; Wilson *et al*., 1999) at Gramene (Ware *et al*., 2002). Also, according to the Cornell RFLP map at Gramene, the *PHYA*- and *PHYC*-containing BACs are included in a region mapped to a quantitative trait locus (QTL) with Gramene identifier GR: 0060322. This QTL is associated with a phenotype of many tillers with fine culmns and normal height. The *PHYB*-containing BAC is also present within a QTL associated with photoperiod sensitivity (Gramene identifier GR: 006067; Causse *et al*., 1994; Wilson *et al*., 1999).

Annotation of the Three PHY*-containing Genome Sequences*

ANNOTATE KNOWN, UNKNOWN, AND PREDICTED GENES

Sequence comparison and gene prediction data was combined to annotate potential genes based on the nature and extent of evidence for their existence. All of the genes annotated within the three assembled BAC sequences, except for the three *PHY*, are considered "potential" because there is no experimental evidence for their existence. The only evidence is from similarities to other sequences or from gene prediction algorithms. In contrast, the *PHY* genes are known to hybridize with well-characterized phytochrome cDNA probes (data not shown).

[12] GenBank accession number AF377946.
[13] GenBank accession number AF461424.

[14] GenBank accession number AF377947.

Potential genes fall into one of three categories. From this point on, these genes will be referred to as *known, unknown* or *predicted* as defined below, with the designation of "potential" being understood.

(1) *Known* genes are those that display a minimum level of sequence identity to other genes of known function. The level of sequence identity necessary to positively identify a gene as known is ultimately a judgment left to the individual annotator. Also, the sequence identity necessary for gene identification or annotation typically varies depending upon the origin of the sequence being used for comparison and the database to which the sequence is being compared. There is no community-wide standard for annotating such genes. For this reason the actual statistics of each comparison result used to determine the annotation of a known gene are reported in the final annotations so that a researcher using the BAC sequences annotated here will know the exact evidence used in the final determination of a gene (Tables 4.6, 4.7 and 4.8).

(2) *Unknown* genes have sequence identity evidence that supports their existence, yet no function can be discerned. Again, the statistical results of the most relevant sequence similarity results are reported in the final annotations (Tables 4.6, 4.7 and 4.8).

(3) *Predicted* genes are those that were identified by multiple prediction programs across more than 80% of their length, yet no other supporting evidence for their existence has been found. The prediction programs that predicted the gene are listed in the final annotations (Tables 4.6, 4.7 and 4.8).

Following curation of the annotation data for all three BACs, it was determined that a total of 55 potential genes and three *PHY* are present (Tables 4.6, 4.7, and 4.8 and Figures 4.2, 4.3 and 4.4). Twenty-three genes, not including *PHY*, were annotated as known and assigned a putative function based on their sequence similarity to genes or protein domains of known function. The gene models depicted in the figures represent the genes from the start of the 5' UTR through the end of the 3' UTR when this information is known. If not, the gene models represent the gene sequence from the +1 ATG to the end of the stop codon. Introns are not represented in the gene models in Figures 4.2, 4.3, and 4.4. Twenty-two genes were identified as unknown based on their sequence similarity to EST and full-length rice cDNA sequences. They could not, however, be assigned a putative function. Of these 22 genes, 5 displayed high sequence identity (E-values less than 2e-64; greater than 44% identity across the entire length of translated sequences) to expressed Arabidopsis genes of unknown function. Ten genes were classified as predicted, having no sequence identity evidence to support their existence. Predicted genes represented in the figures are indicative of sequences from the +1 ATG to the end of the start codons. These models do not contain UTR information.

A total of 466 kb of finished sequence was obtained from these three BACs, yielding gene densities for OSJNBa0031O09 (*PHYA*), OSJNBa0016B07 (*PHYB*), and OSJNBa0032E21 (*PHYC*) of one per 10.4, 10.8, and 8.1 kb, respectively, when only known and unknown genes were included in the calculation. These densities would obviously be somewhat higher if predicted genes were included.

SEARCH FOR ANCIENT TRANSPOSONS AND PHYTOCHROMES

The three BAC sequences were compared to the McCarthy/McDonald collection of LTR retrotransposon, transposon, and transposase sequences. Only one sequence was identified as

matching to the BAC sequences with an E-value of less than 1.0 and percent identity (%ID) greater than 30% across the entire length of the sequence. The *PHYC*-containing BAC OSJNB0032E21 contains an area of sequence that matches a TAM-1 like transposase (GenBank accession number AAK00419.2; Table 4.8 and Figure 4.4, Feature Number 17.)

To determine whether any ancient phytochromes were present in the BAC sequences the *PHY*-containing BAC sequences were also compared to the *PHY* sequences listed in table 4.3. The results that were returned from these comparisons all had E-values greater than 1.0 and none matched with greater than 10% sequence identity across the entire length of the *PHY* genes. In short, no fragments of *PHY* or or low level sequence homology to *PHY* could be detected within the BAC sequences, other than the full-length functional *PHY* (phytochromes A, B, and C) already known to be present.

Comparing PHY*-containing Genomic Sequences*

Each *PHY*-containing rice BAC was compared to the other two with the five methods described previously. Except for the sequence similarity between the *PHY* themselves, there were no regions surrounding the phytochrome genes of sequence conservation longer than 100 nucleotides detected by any of these methods (data not shown). Furthermore, when the BAC sequences were screened by RepeatMasker (Smit, Hubley, and Green, 1996) prior to comparison greater than 90% of the total number of regions of sequence similarity previously detected were no longer present, indicating that most of the sequence similarity being reported was due to simple repeat sequences.

The *PHY*-containing rice BACs were also compared to other *PHY*-containing genomic sequences from Arabidopsis and sorghum. Eighteen comparisons were performed using the five

methods described above. One of these comparisons, namely that between *PHYA*-containing sequences from sorghum and rice, had already been performed and published (Morishige *et al*., 2002). This comparison served as a positive control with which to verify the robustness of the five comparison methods presented here. Like the intra-rice sequence comparisons, all of the rice-Arabidopsis and rice-sorghum comparisons, except that of the *PHYA*-containing rice/sorghum sequences, yielded no regions of sequence conservation detected that were longer than 100 nucleotides other than those between the *PHY* (data not shown). My own comparison of the *PHYA*-containing genomic sequences from sorghum and rice corresponded to that published previously, indicating that the comparison methods presented here are comparably robust. These results indicate that the 120-200 million years that separate the paralogous regions of the *PHY* family in rice, as well as between the rice and *Arabidopsis* orthologs and paralogs, are too many to permit the detection of any sequence or gene conservation by any of these methods, if in fact such areas of conservation ever existed. Therefore, while upstream and downstream portions of the genome may have been included in the gene/genome duplication that led to a three-member monocot *PHY* family, it appears that too much evolutionary time has passed to detect any mechanism of *PHY* family expansion.

Discussion

Employing Multiple Gene Prediction Programs Improves Annotation

It is well documented that prediction programs tend to have a high false-positive rate and do not usually predict alternatively spliced variants accurately. Combining results from multiple prediction programs, however, will increase prediction accuracy in genome annotations (Shah *et al*., 2003).

The final annotation of the three BACs contained 48 genes that displayed some sequence identity to previously sequenced genes when the BAC sequences were compared to the sequence databases listed in Table 4.2. Figures 4.5, 4.6, and 4.7 are screen shots of the Artemis displays containing all of the gene predictions (Figures 4.5, 4.5, and 4.7 tracks A-D) as well as results from five of the sequence comparisons described in Table 4.2 (tracks E-I). The final 48 genes were annotated by combining the gene prediction and sequence comparison data to determine the most likely gene model. These 48 genes were audited to determine which gene prediction program, if any, completely or partially predicted their occurrence. FGeneSH was by far the most accurate prediction program of those used here, predicting 27 of the 48 genes exactly and partially predicting 9 of the 48 genes. GeneMark, Genscan, and RiceHMM correctly predicted 13, 4, and 2 of the 48 genes exactly. Although RiceHMM was the least successful in predicting the exact gene model corresponding to the full-length cDNA sequence data, the 4% success rate is somewhat misleading. Almost 50% of the predictions generated by RiceHMM correctly predicted at least 80% of the exons included in the final annotated gene model.

The Significance of a Comparison Result is Directly Related to the Sequence Being Used for Comparison and the Database to Which it is Compared

There is one fundamental question regarding the task of assigning identities or annotating based on sequence similarity evidence: which sequence similarity results are "significant" and which are not?

Currently there is no community-wide standard for utilizing sequence comparison information to annotate genes within a nucleotide sequence of unknown gene content. For this reason, the statistics and sequence similarity information regarding each piece of evidence used

to determine the final annotation of a gene within the BAC sequences was included in the annotation record.

Also, the type of sequence used in the comparison and the database to which the sequence is compared directly affects the standards by which the comparison results need to be evaluated. To highlight this point, two global observations that were made during the course of annotating the three assembled BAC sequences are described below.

(1) Comparisons of the BAC sequences to rice full-length cDNAs result in several HSPs with high %ID and low E-values.

In general, a cDNA sequence aligns to contiguous, yet interrupted regions of the BAC genome sequence, resulting in the alignment being reported as multiple high sequence pairs (HSPs). Furthermore, the results will yield statistical scores that are dependent upon the type of sequence hit by the subject, as well as the species from which it comes. For example, when comparing rice genomic DNA to rice full-length cDNAs downloaded from KOME, the percentage identity (% ID) and E-values are expected to be very high and low, respectively, as both the query and subject sequence are high quality rice sequences and therefore should match almost perfectly. The match, however, will be interrupted and distributed among multiple HSPs. Therefore, only those matches with E-values less than 1e-80 and % ID values greater than 80% across the entire length of the match were reported in the annotation records (Tables 4.6, 4.7, and 4.8). The final % ID values reported were calculated by adding the total number of matching nucleotides across the entire length of all HSPs within an hit and dividing by the total number of nucleotides present in all HSPs. The following formula further describes the method by which these final %ID values were calculated: $((\%ID_{HSP\ 1}/100 * L_{HSP\ 1}) + (\%ID_{HSP\ 2}/100 * L_{HSP\ 2}) + \ldots$

+ ($\%ID_{HSP\ N}$ /100 * $L_{HSP\ N}$) / ($L_{HSP\ 1} + L_{HSP\ 2} + \ldots + L_{HSP\ N}$)). These calculated % ID values are those reported in the final annotation when a hit to a cDNA was included as evidence in the annotation record for the three BACs annotated here.

(2) Comparisons of the BAC sequences to EST databases result in HSPs with lower %ID and higher E-values than when compared to rice full-length cDNAs.

Alternatively, when nucleotide sequences from the BACs were compared to the est_others database the comparison results contained sequences from many species other than rice. Also, EST sequences in this database are typically shorter and of more dubious quality than the rice full-length cDNAs from KOME. The statistical values of the comparison results reflect this inherent limitation of this database. Therefore, comparison results with E values less than 1 e-30 were in this instance considered significant evidence of the existence of a gene.

Genes Surrounding PHY *Are Not Functionally Related*

To determine if the genes surrounding *PHY* were also involved in photomorphogenic responses, or in signaling networks that are known to include *PHY*, all known genes were classified by their molecular function as defined through the gene ontology (Ashburner *et al*., 2000). The genes surrounding the phytochromes seem to encode proteins that function in all areas of plant development, including signal transduction, DNA recombination, replication and repair, and lipid, inorganic ion, carbohydrate, and amino acid transport and metabolism, without leaning towards a particular functional classification (Figure 4.8).

For example, genes surrounding *PHYA* on OSJNBa0031O09 (Table 4.6 and Figure 4.2) include those that are involved in amino acid, lipid, and carbohydrate metabolism and transport,

as well as a transcriptional regulator and two genes involved in signal transduction. The transcriptional regulator gene product was identified based on its sequence similarity to *Arabidopsis* protein NP_176659, which was in turn identified based on its similarity to human HsGCN1 (accession number AAC51648). It has been hypothesized that GCN1 regulates GCN2 kinase activity (Marton *et al*., 1997). GCN2 is a transcriptional activator of GCN4, which itself is a transcriptional activator of amino-acid biosynthetic genes in *Saccharomyces cerevisiae* (Wek *et al*., 1990). The two genes involved in signal transduction immediately flank either side of *PHYA* and are oriented in the same direction as the phytochrome gene. The 5' flanking gene, a putative protein kinase, was identified based on its sequence similarity to an *Arabidopsis* gene that has been described as a member of the protein kinase family. This function, however, has not been experimentally verified. Protein kinases, which are organized into signaling cascades, are important components of cellular regulatory systems (Wang *et al*., 2003). The 3' flanking gene was highly similar to *Arabidopsis* serine/threonine kinase P43293. This sequence, isolated by low stringency hybridization with the catalytic domain of a putative plant receptor protein kinase, has been suggested as playing a role in the regulation of plant growth and development (Moran and Walker, 1993).

The *PHYB*-containing BAC (OSJNBa0016B07) contains genes that encode proteins known to be involved in lipid and carbohydrate metabolism, as well as one tam-3 like transposase (Table 4.7 and Figure 4.3). There are, however, no other genes within OSJNBa0016B07 involved in signal transduction, photomorphogenesis, or translational activation identified at this time.

OSJNBa0032E21 (the *PHYC*-containing BAC) contains the highest number of genes whose protein products are involved in either signal transduction or translational activation

(Table 4.8 and Figure 4.4). Two TypeII MADS-box proteins are located in tandem ~ 40,000 bp 3' from *PHYC (*Fig. 4.4). MADS-box genes are crucial components in the regulation of root, flower, seed, and fruit development (Ng and Yanofsky, 2001). In plants, Type II MADs proteins contain a variable C-terminal domain known to confer specificity to the individual protein (Lamb and Irish, 2003). They are involved in the formation of protein complexes and transcriptional activation (Riechmann and Meyerowitz, 1997; Egea-Cortines, *et al*., 1999; Honma and Goto, 2001). A putative oligopeptide transporter is located upstream, or 5', of *PHYC*. Although the exact role of plant peptide transport is still largely undefined, the *Arabidopsis* genome sequence contains ten times more predicted peptide transporters than any other sequenced organism, prokaryotic or eukaryotic (Stacey *et al*., 2002). The abundance of these peptide transporters has led some to suggest that peptide transporters play a diverse and important role in plant development and growth (Stacey *et al*., 2002). A putative topoisomerase is also located immediately 3' of *PHYC* (Figure 4.4). Topoisomerases are able to pass one DNA duplex through another to solve various topological problems that arise from the processing of double-helical DNA. An example of this function is the removal of the superhelical twist generated from DNA replication and transcription, which can lead to perturbed gene expression if not appropriately dealt with (Liu and Wang, 1987; see Corbett and Berger, 2003, for minireview). The *PHY*C-containing BAC also has genes present whose products encode proteins involved in inorganic ion transport and structural cell wall components.

While no obvious direct functional link can be made between *PHY* and its surrounding genes, *PHY* interactions and functions in diverse signaling networks are still largely undefined, as are the exact functions of many of the genes identified here.

More Sequences are Necessary to Investigate the Mechanisms of PHY *Family Expansion*

The mechanism of gene duplication inarguably imparts an opportunity for the increased fitness of an organism by increasing its genetic diversity (Lynch *et al*., 2000). Indeed, a recent analysis of the *Arabidopsis thaliana* genome has revealed that most of it consists of paralogous genes, or genes that arose from a common ancestor (Raes *et al*., 2003). It has been further hypothesized that these genes most likely originated through one or more genome duplication events that occurred at some unknown point in the ancient past of *Arabidopsis thaliana* (Raes, *et al*., 2003).

Although it is generally understood that *PHY* complexity was achieved through gene or genome duplications followed by divergence, the mechanisms by which the *PHY* genes duplicated are unclear. Analyses have shown that large-scale genomic duplications have occurred in a number of organisms (Gale and Devos, 1998). Indeed, not only do individual genes duplicate, but blocks of multiple genes or even entire chromosomes or genomes are now thought to have contributed to the evolution of animals, fungi, and plants. These duplications result in regions of similar gene content and order within an organism. Yet, genomes can vary tremendously in size and organization, even among closely related organisms. This variation seems to be mainly the result of recombination events, horizontal gene transfer, transposon activity, gene duplication, and gene loss (Bancroft, 2000). These events can make it very difficult to find statistically relevant homologous regions within the genome, particular when the duplication events are very old.

It was initially hypothesized that comparing genome regions flanking the three *PHY* could possibly indicate how much of the surrounding DNA was duplicated with the *PHY* as the family expanded. The *PHY* family in monocots is thought to have evolved from a single cyanobacterial progenitor by a minimum of two duplication and divergence events

approximately 200 Mya (Wolfe *et al.*, 1989; Donoghue and Mathews, 1999). The present study revealed that, using the methods employed here, no regions of sequence similarity larger than 100 nucleotides upstream and downstream of *PHY* could be identified. Moreover, 90% of those regions of similarity of less than 100 nt in length were simple repeat sequences. Given the presumed ancient evolutionary history of the putative duplication events that resulted in the three-membered *PHY* family, however, this observation of little to no detectable sequence similarity between *PHY*-containing genomic sequences of rice is certainly reasonable.

The ability to apply information gleaned from comparative genome projects as well as comparative mapping studies across the monocot/eudicot divide has been much debated (Salse *et al*. 2003). *PHY*-containing regions of Arabidopsis and rice were compared to investigate this phenomenon within these regions of these two plant genomes. All of the comparisons undertaken in this study revealed little to no sequence conservation except for that between *PHY* when *PHY*-containing sequences from rice and Arabidopsis were compared.

Overall, with the exception of the *PHYA*-containing rice/sorghum comparisons (previously published by Morishige *et al*., 2002, and confirmed here), none of the comparisons yielded any detectable regions of sequence conservation. The *PHYA*-containing rice BAC OSJNBa0031O09 sequenced in this study did reveal high degrees of gene conservation downstream of *PHYA* in rice, sorghum, and maize (Morishige *et al.*, 2002), suggesting that the duplication of the *PHYA* progenitor did not happen independently in rice, but rather before these plants diverged. Nine open reading frames were identified in common between the 100-kb *PHYA*-containing genomic region when comparing sorghum and rice. The ortholog of one gene, dihydroxybutanone kinase (DHBK) was missing on the rice BAC, suggesting that some deletion, duplication, or translocation event had occurred in the region (Morishige *et al*., 2002).

Given that comparison of the orthologous *PHYA* regions of rice and sorghum yielded detectable regions of conserved gene order it is probable that sequence conservation would be detected surrounding *PHY* if these genome sequences were compared to other orthologous *PHY*-containing DNA from a plant such as sorghum, whose genome shared a common ancestor with rice more recently than the 200 million years of evolution that separates the rice genome and Arabidopsis. The production of more genomic sequences from the grass family is necessary to investigate the presence or absence of colinearity within this region of the genome as compared to its closer plant relatives. Indeed, many groups are currently working on targeted genome sequencing from other species such as maize (Barbazuk *et al*., 2003) and sorghum (http://pgir.rutgers.SorghumViewer/BACsorghum_ viewer.html). Moreover, the production of genome sequence information from any plant species, not just the grasses, will greatly improve the fate of comparative genomics in plants. The inception of genome projects in tomato (http://www.sgn.cornell.edu/about/tomato_project/) and potato (http://www.potatogenome.org/nsf3/goals_research) will undoubtedly help with comparative genomics studies within the plant kingdom.

Conclusions

This study sought to (I) sequence the *PHY* family and flanking genomic sequences in rice, (II) extensively and accurately annotate these genomic sequences, and (III) compare these genomic sequences of rice to each other and to other plant *PHY*-containing sequences to investigate possible mechanisms of *PHY* family expansion.

The *PHY* family and the genomic sequence surrounding each member was successfully sequenced by a shotgun approach. These BAC sequences were confirmed by restriction digest analysis and deposited into GenBank for availability to the genome community at large.

The three BAC sequences were extensively annotated by manually curating gene prediction and sequence similarity data generated by running multiple prediction algorithms and performing many comparisons to databases of known genes, proteins, and protein domains. It was determined that employing multiple prediction programs greatly increases the overall accuracy of gene prediction and that, of the prediction algorithms used here, FgeneSH was the most accurate. The genes surrounding *PHY* were found not to display any discernable functional relationship to one another.

These *PHY*-containing genomic sequences of rice were compared to each other and to other plant *PHY*-containing sequences to investigate any possible mechanisms of *PHY* family expansion in plants. Although 21 comparisons were performed with five different methods, no evidence or information regarding the mechanisms of *PHY*-family expansion could be uncovered. The methods utilized for these comparisons, however, were confirmed to be robust and the lack of detectable sequence similarity within these areas of the rice, Arabidopsis, and sorghum genomes is not surprising given the ancient nature of the phytochrome gene duplications.

Figures and Tables

Table 4.1: Files of FastA Sequences Prepared for Database Comparisons. The following files were generated in order to compare the sequences within them to databases of genes or protein sequences or other known sequence repositories. The sequences were generated either with commands present in the Artemis file menu or command options in GCG.

File Number	File Type	Description of sequence(s) within each file
1	Nucleotide	Sequence from each predicted exon within an assembled BAC sequence
2	Amino Acid	Sequence from each predicted exon within an assembled BAC sequence
3	Nucleotide	Sequence from all ORFs longer than 50 amino acids within an assembled BAC sequence
4	Amino Acid	Sequence from all ORFs longer than 50 amino acids within an assembled BAC sequence
5	Nucleotide	The assembled BAC sequence divided into 3-kb overlapping pieces*
6	Nucleotide	The whole BAC sequence**

*Sequences generated with “breakup” command in GCG.
**Sequence generated with “tofasta” command in GCG.
The sequences in file numbers 1-4 were generated in Artemis.

Table 4.2: Databases to Which the Six Files of Nucleotide and Amino Acid Sequences Listed in Table 4.1 Were Compared.

File/Sequence Type	Databases to Which the Sequences Were Compared
Nucleotide sequence of all predicted exons	KOME, NR, est_others
Amino acid sequence of all predicted exons	PIR_NREF, swiss prot plus, Pfam
Nucleotide sequence of all ORFs > 50 amino acids in length	KOME, NR, est_others
Amino acid sequence of all ORFs > 50 amino acids in length	PIR_NREF, swiss prot plus, Pfam
3-kb overlapping nucleotide sequences derived from entire BAC sequence	KOME, NR, est_others
Whole BAC sequence	KOME, NR, est_others, htgs, swiss prot plus

Table 4.3: *PHY* Sequences Retrieved from GenBank.

Organism	Phytochrome	GenBank Accession Number
Oryza sativa	A	AB109891
Oryza sativa	B	AB183525
Oryza sativa	C	AB141942
Sorghum bicolor	B	AF182394
Sorghum bicolor	C	AAR33022
Arabidopsis thaliana	A	NM_100828
Arabidopsis thaliana	B	AY466946
Arabidopsis thaliana	C	AY394847
Arabidopsis thaliana	D	NM_117721
Arabidopsis thaliana	E	NM_117923

Table 4.4: *PHY*-containing Genomic Sequences Used in Comparative Analyses.

Sequence Name	GenBank accession number	Description	Journal Reference
OSJNBa0031O09	AF377946	*PHYA*-containing BAC from rice	this thesis
OSJNBa0016B07	AF461424	*PHYB*-containing BAC from rice	this thesis
OSJNBa0032E21	AF377947	*PHYC*-containing BAC from rice	this thesis
T22J18	AC003979	*PHYA*-containing BAC from *Arabidopsis*	unpublished
MSF3	AC005724	*PHYB*-containing BAC from *Arabidopsis*	unpublished
Chromosome 5, complete sequence	NC_003076	*PHYC*-containing BAC from *Arabidopsis*	unpublished
AL161543	AL161543	*PHYD*-containing BAC from *Arabidopsis*	unpublished
AL161548	AL161548	*PHYE*-containing BAC from *Arabidopsis*	unpublished
Sbb3766	AF369906	*PHYA*-containing BAC from *Sorghum bicolor*	Morishige *et al*., 2002

Table 4.5: Comparisons of *PHY*-containing Genomic Sequences Used to Investigate Possible Mechanisms of *PHY* Family Expansion Within the Rice Genome and Between Rice and Arabidopsis.

Comparison Name	Sequence 1	Sequence 2
Os[1]Os; *PHYA* versus *PHYB*	OSJNBa0031O09	OSJNBa0016B07
OsOs; *PHYA* versus *PHYC*	OSJNBa0031O09	OSJNBa0032E21
OsOs; *PHYB* versus *PHYC*	OSJNBa0016B07	OSJNBa0032E21
OsAt[2]: *PHYA* versus *PHYA*	OSJNBa0031O09	T22J18
OsAt: *PHYA* versus *PHYB*	OSJNBa0031O09	MSF3
OsAt: *PHYA* versus *PHYC*	OSJNBa0031O09	Chromosome 5, complete sequence
OsAt: *PHYA* versus *PHYD*	OSJNBa0031O09	AL161543
OsAt: *PHYA* versus *PHYE*	OSJNBa0031O09	AL161548
OsSb[3] : *PHYA* versus *PHYA*	OSJNBa0031O09	Sbb3766
OsAt: *PHYB* versus *PHYA*	OSJNBa0016B07	T22J18
OsAt: *PHYB* versus *PHYB*	OSJNBa0016B07	MSF3
OsAt: *PHYB*versus *PHYC*	OSJNBa0016B07	Chromosome 5, complete sequence
OsAt: *PHYB*versus *PHYD*	OSJNBa0016B07	AL161543
OsAt: *PHYB* versus *PHYE*	OSJNBa0016B07	AL161548
OsSb : *PHYB* versus *PHYA*	OSJNBa0016B07	Sbb3766
OsAt: *PHYC* versus *PHYA*	OSJNBa0032E21	T22J18
OsAt: *PHYC* versus *PHYB*	OSJNBa0032E21	MSF3
OsAt: *PHYC*versus *PHYC*	OSJNBa0032E21	Chromosome 5, complete sequence
OsAt: *PHYC*versus *PHYD*	OSJNBa0032E21	AL161543
OsAt: *PHYC* versus *PHYE*	OSJNBa0032E21	AL161548
OsSb : *PHYC*versus *PHYA*	OSJNBa0032E21	Sbb3766

1: Os = *Oryza sativa*
2: At = *Arabidopsis thaliana*
3: Sb = *Sorghum bicolor*

Table 4.6: List of Annotated Features of OSJNBa0031O09.

BAC Name: OSJNBa0031O09
PHYTOCHROME A

Feature Number	Feature Type	Function	Sequence Identity Evidence
1	Putative Transcriptional Activator with adaptin domain	Translation, ribosomal structure, biogenesis	E=0.0, %ID=100 to Rice full length cDNA sequence CLUSTER_ID=727; E=0.0, %ID=66 to *Arabidopsis* translational activator NP_176659; E=2e-170 and 99.8% aligned to KOG1242 protein containing adaptin N-terminal region
2	Unknown Gene	Unknown	E=1e-164, %ID=52% across entire length to NP_199208 which is an expressed *Arabidopsis* gene of unknown function
3	Unknown Gene	Unknown	E=1e-118, %ID=92% to *Hordeum vulgare* subsp. vulgare cDNA clone HX05I09 5-PRIME; E=2e-41 with 85.5% alignment to kog3312 (predicted membrane function); E=1e-54 and %ID=60 to *Arabidopsis* AAK63951
4	Unknown Gene	Unknown	E=0.0, %ID=96.9 to Rice full length cDNA sequence CLUSTER_ID= 16453
5	Putative Histone H2A	DNA binding	E=0.0, %ID=98.5 to Rice full length cDNA sequence CLUSTER_ID= 12879; E=3e-58; %ID=86% to Histone H2A; E=7e-32 and 91.3% aligned to cd00074 Histone H2A domain
6	Unknown Gene	Unknown	E=0.0, %ID=100 to Rice full length cDNA sequence CLUSTER_ID= 5060; E=2e-64, %ID=44 to NP_197604 (expressed protein of unknown function in *Arabidopsis*)
7	Predicted Gene	Unknown	None
8	Unknown Gene	Unknown	E=0.0, %ID=90.7 to Rice full length cDNA sequence CLUSTER_ID=13020
9	Unknown Gene	Unknown	E=0.0, %ID=99 to Rice full length cDNA sequence CLUSTER_ID=1679;
10	Unknown Gene	Unknown	E=0.0, %ID=94.8 to Rice full length cDNA sequence CLUSTER_ID=2905
11	Predicted Gene	Unknown	None
12	Putative glutamate	Amino acid	E=1e-106, %ID=90.4 to Rice full length cDNA sequence

	decarboxylase	transport and metabolism	CLUSTER_ID=5981; E=0, %ID=79, % POS=91 to *Arabidopsis* GAD1
13	Putative 3.4 dihydroxy 2 butanone kinase	Carbohydrate transport and metabolism	E=0.0, %ID=62, %POS=71 to 3,4-DIHYDROXY-2-BUTANONE KINASE in tomato
14	Putative lysophospholipase	Lipid transport and metabolism	E=0.0, %ID=100 to Rice full length cDNA sequence CLUSTER_ID=20807; lysophospholipase domain detected in Pfam
15	Putative protein kinase	Signal transduction mechanisms	E=0.0, %ID=100 to Rice full length cDNA sequence CLUSTER_ID=16190; E=2e-55, %ID=37, %POS=55 to human protein kinase
16	*PHYA*	Photomorphogenesis	E=0.0, %ID=100 to Rice full length cDNA sequence CLUSTER_ID=1032
17	Putative serine/threonine protein kinase	Signal transduction mechanisms	E=2e-55 to *Arabidopsis* serine/threonine kinase P43293.
18	Putative oligopeptide symporter	Amino acid transport and metabolism	E=0.0, %ID=100 to Rice full length cDNA sequence CLUSTER_ID=2992; E=0.0, %ID=69 to *Arabidopsis* PT2B

Table 4.7: List of Annotated Features of OSJNBa0016B07.

BAC Name: OSJNBa0016B07
PHYTOCHROME B

Feature Number	Feature Type	Function	Sequence Identity Evidence
1	Putative O-linked GlcNAc	Carbohydrate transport and metabolism	E=0.0, %ID=100 to rice full length CLUSTER_ID=2113; E=0.0, %ID=63 to O-linked GlcNAc transferase from *Arabidopsis* BAA94982
2	Unknown Gene	Unknown	E=0.0, %ID=100 to rice full length CLUSTER_ID= 12780; 100% aligned to Fip1 motif, E=9e-13 to pfam05182
3	Unknown Gene	Unknown	E=0.0, % ID=100 to rice full-length cluster_id 5256. Also hits E=e-120, %ID=58 to *Arabidopsis* expressed protein of unknown function NP_191796
4	Unknown Gene	Unknown	E=0, %ID=92 to rice full-length CLUSTER_ID=17942
5	*PHYB*	Photomorphogenesis ; light-regulated signal transduction	100% aligned to pfam 00360 phytochrome domain E=4e-96; E=0, %ID=100 to acc_num BAC76432 (rice phyb)
6	Predicted Gene	Unknown	None
7	Putative pectinesterase	Carbohydrate transport and metabolism	96% aligned to pfam01095 (pectinesterase) E=2e-48; E=0, %ID=100 to rice CLUSTER_ID=13954; E=e-100. %ID=52 to pectinesterase family from *Arabidopsis* NP_188331
8	Predicted Gene	Unknown	None
9	Putative PPR2	Chloroplast localized protein product necessary for plastid ribosome accumulation	E=0.0, %ID=74 to *Zea mays* PPR2 (accession number=AAP37977);
10	Predicted Gene	Unknown	None
11	Putative GSDL-motif lipase hydrolase	Predicted to be involved in lipid metabolism	E=3e-98 %ID=51 to GDSL-motif lipase hydrolase NP_190609 from *Arabidopsis*. E=0.0, %ID=100 to rice full length CLUSTER_ID=8795; E=7e-17, %ID=89.2 to cog3240 and pfam pf00657

12	Putative Enoyl-CoA hydratase	Lipid transport and metabolism	E=0.0, %ID=100 to CLUSTER_ID=3515; 90% aligned to kog1680 (enoyl-coa hydratase, E=3e-53); E=2e-99 and %ID=67 to putative enoyl coa hydratase from *Cicer arietinum* (chickpea), accession number = CAB1740
13	Unknown Gene	Unknown	E=e-175, %ID=71% to expressed protein of unknown function in *Arabidopsis* NP_193352. E=0.0, %ID=100 to rice full length CLUSTER_ID=3398.
14	Unknown Gene	Unknown	E=2e-65, %ID=91.3 to pfam03140, a plant protein of unknown function;; E=0, %ID=100 to rice full length CLUSTER_ID=18920
15	Unknown Gene	Unknown	E=0.0, %ID=100 to rice full length CLUSTER_ID=7298; an internal ORF hits AAD472313, a hypersensitive reaction associated CA2+ binding protein from a bean with E=8e-25, %ID= 47, and %POS=56
16	Putative transposase	Replication, recombination and repair	E=1e-117, %ID =86 to *Oryza sativa* putative transposase NP921575; E=2e-19 to kog 1121, a tam3-transposase
17	Putative haloacid dehalogenase-like protein		E=0, %ID=100 to rice full length CLUSTER_ID=577; E=0.0, %ID=60 to haloacid dehalogenase-like hydrolase family from *Arabidopsis* NP_564718; E=3e-19, %ID=100 to pfam00702, a haloacid dehalogenase-like hydrolase
18	Predicted Gene	Unknown	None
19	Unknown protein	Unknown	E=0.0, %ID=100 to rice full length CLUSTER_ID=15465; E=2e-19, %ID=43, %POS=59 to bHLH protein from *Arabidopsis* (accession number) NP_563839
20	Predicted Gene	Unknown	None
21	Predicted Gene	Unknown	None

Table 4.8: List of Annotated Features of OSJNBa0032E21.

BAC Name: OSJNBa0032E21
PHYTOCHROME C

Feature Number	Feature Type	Function	Sequence Identity Evidence
1	Unknown Gene	Unknown	E=0.0 and %ID=100 to rice full length CLUSTER_ID=1283.
2	Predicted Gene	Unknown	None
3	Unknown Gene	Unknown	E=0.0 and %ID=87 to rice full length CLUSTER_ID 10993.
4	Unknown Gene	Unknown	E=0.0 and %ID=100 to rice full length CLUSTER_ID 6890. Shows similarity to PIR_NREF protein domain of unknown function.
5	Unknown Gene	Unknown	E=0.0 and %ID=92 to rice full length CLUSTER_ID=3686. Shows similarity to PIR_NREF protein domain of unknown function.
6	Predicted Gene	Unknown	None
7	Putative MADS box	DNA binding, transcriptional regulation	E=2e-29 and 100% aligned to cd00265 MADS box TypeII subfamily of eukaryotic transcriptional regulators; E=2e-20 and 100% aligned to pfam01486 K-box regions commonly found with SRF-type transcription factors. E=e-111 and %ID=100 to MADS-box like protein BAA81882; E=0 and %ID=99 to rice full length CLUSTER_ID=13688.
8	Putative AP1-like MADS box	DNA binding transcriptional regulation	E=1e-23 and 100% aligned with MADS box TypeII subfamily; E=e-108 and %ID=99 aligned to BAA943432 AP1-like MADS box protein.
9	Unknown Gene	Unknown	E=0.0 and %ID=97 to rice full length CLUSTER_ID=21437. Shows 55% identity to 200 amino acids of an expressed *Arabidopsis* gene of unknown function.
10	Putative cysteine proteinase	Post-translation modification, protein turnover, chaperones	E=e-122 and %ID=74 to barley cysteine proteinase T06207. E=1e-58 and 91.4% aligned to kog1543 cysteine proteinase domain
11	Putative potassium outward rectifying channel	Inorganic ion transport and metabolism	E=0.0 and %ID=99% to rice full length CLUSTER_ID=4511. E=5e-17 and 51% aligned to tandem pore domain K+ channel KOG1418; E=e-110 and %ID= 57 to eucalyptus outward-rectifying potassium channel.

12	Putative topoisomerase	DNA replication, recombination, and repair	E=5e-34 and %ID=76 to rice putative topoisomerase AAP68363.
13	*PHYC*	Photomorphogenesis	E=0.0 and %ID=100 to Q9ZWI9 rice PHYC protein; E=e-100 %ID=99 to rice full length CLUSTER_ID=8669.
14	Unknown Gene	Unknown	E=0.0 and %ID=100 to rice full length CLUSTER_ID=2369.
15	Unknown Gene	Unknown	E=0.0 and %ID=100 to rice full length CLUSTER_ID=9289.
16	Unknown Gene	Unknown	E=0.0 and %ID=100 to rice full length CLUSTER_ID=3303.
17	Putative TAM-1 like transposase	None	Very similar to putative TAM1 transposon protein TNP2 (O. sativa) acc_num AAK00419.2; E=1e-236 and %ID=98.
18	Putative hydroxyproline rich glycoprotein	None: Structural component of the plant cell wall	Very similar to putative hydroxyproline-rich glycoprotein. E=0.0 to acc_num BAB86566.1 (AP003710) across the entire length of the protein."
19	Putative oligopeptide transporter protein	Signal transduction	E=0.0 and %ID=97 to rice full length CLUSTER_ID=6383; E=6e-156 and %ID=99.4 to pfam03169, an oligopeptide transporter protein.

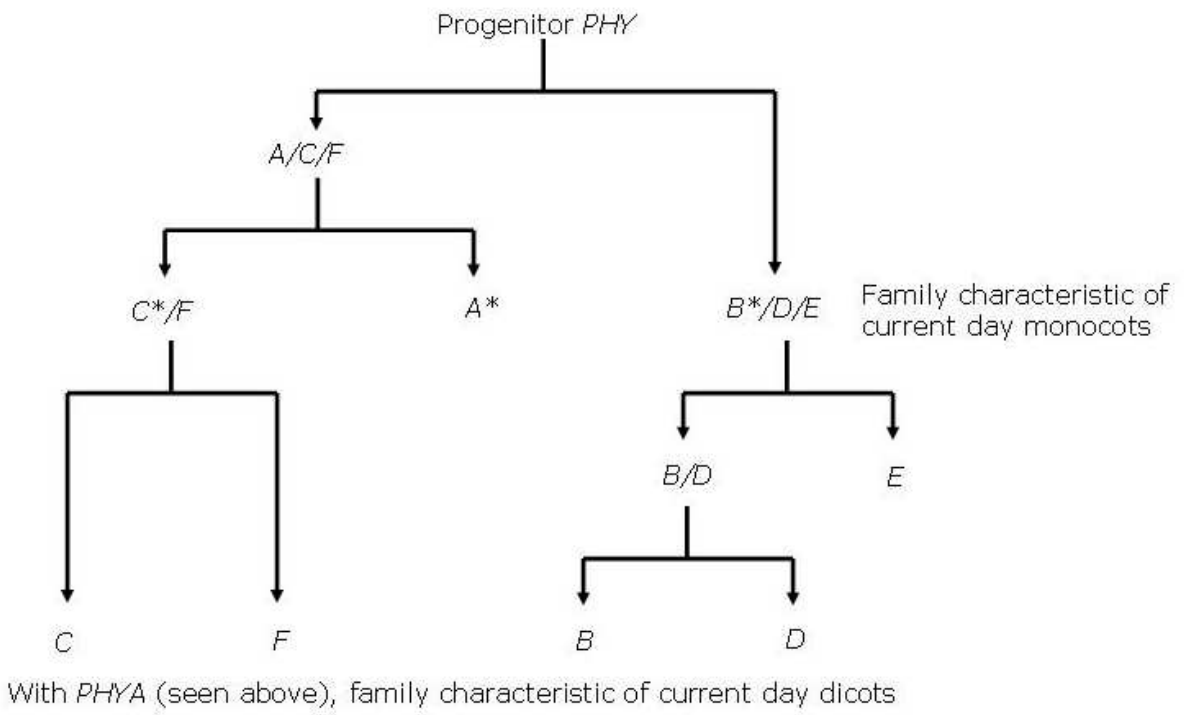


Figure 4.1: *PHY* Evolution in Monocots and Dicots. A progenitor *PHY* duplicated to yield the *PHYA/C/F* and *PHYB/D/E* subfamilies prior to gymnosperm formation. Prior to monocot/dicot formation *PHYA/C/F* underwent a second duplication to form the *PHYA* and *PHYC/F* subfamilies. More recent duplications are thought to have given rise in dicots to *PHYC*, *F*, *B*, *D*, and *E* as illustrated.

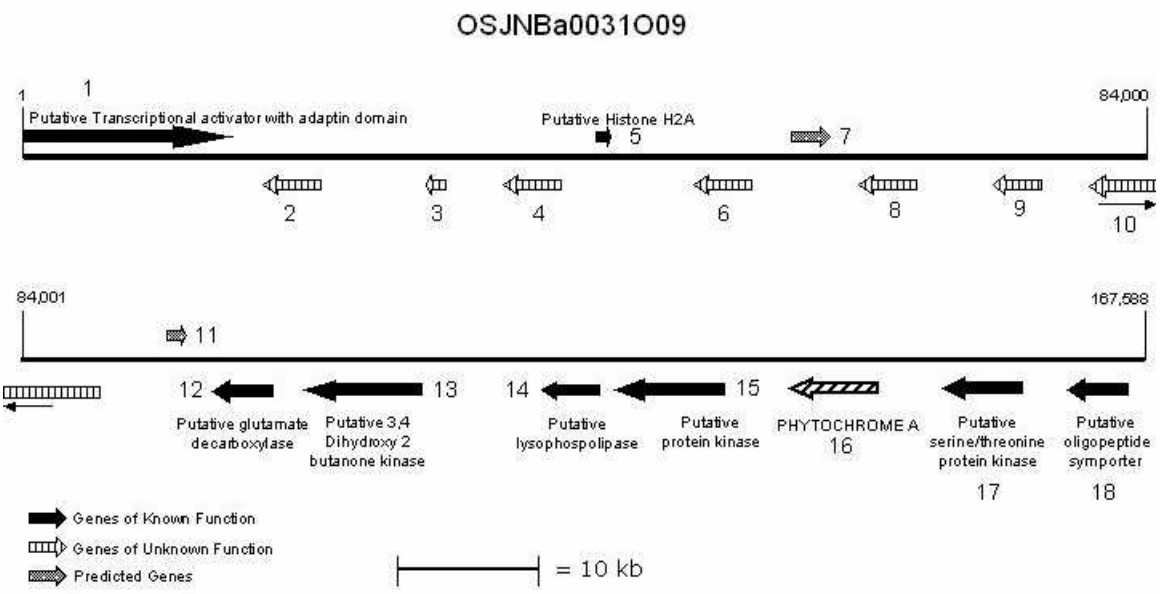


Figure 4.2: Diagram to Scale of all Genes Annotated in BAC OSJNBa0031O09. The gene models depicted here are from the +1 ATG site to the stop codon if they are predicted genes. The models of genes encoding proteins of unknown and known function include 3' and 5' UTRs when known. None of the models depict introns present in the genes. Gene models are identified by the feature numbers in Table 4.6.

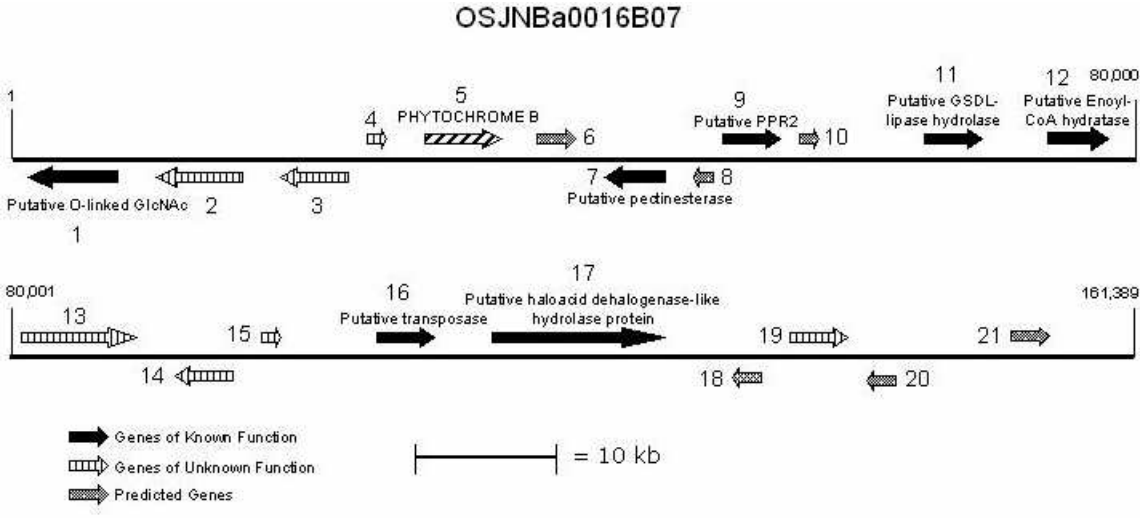


Figure 4.3: Diagram to Scale of all Genes Annotated in OSJNBa0016B07. The gene models depicted here are from the +1 ATG site to the stop codon if they are predicted genes. The models of genes encoding proteins of unknown and known function include 3' and 5' UTRs when known. None of the models depict introns present in the genes. Gene models are identified by the feature numbers in Table 4.7.

.

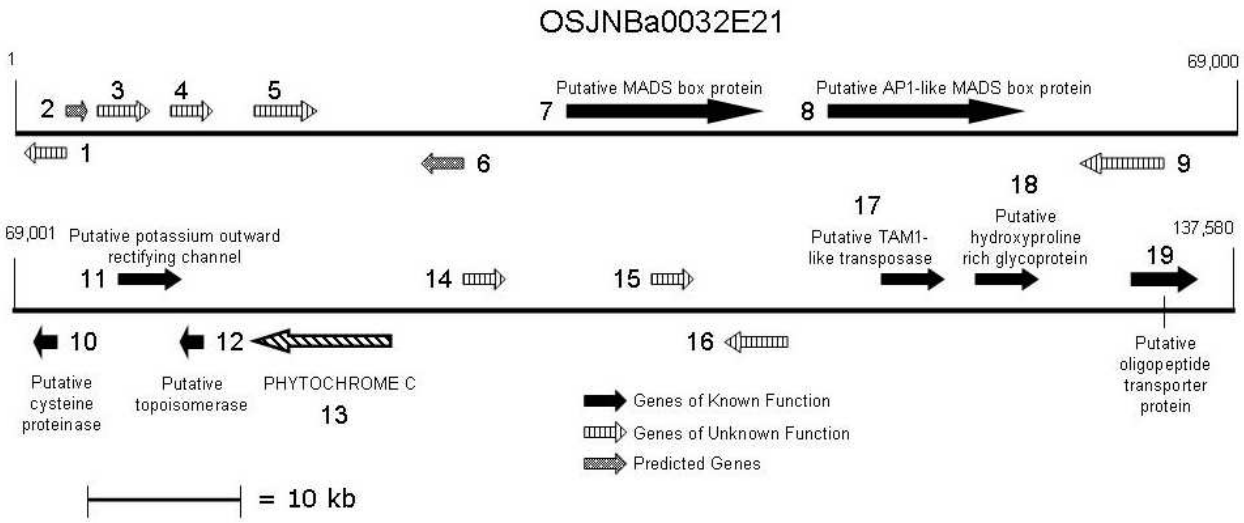


Figure 4.4: Diagram to Scale of all Genes Annotated in BAC OSJNBa0032E21. The gene models depicted here are from the +1 ATG site to the stop codon if they are predicted genes. The models of genes encoding proteins of unknown and known function include 3' and 5' UTRs when known. None of the models depict introns present in the genes. Gene models are identified by the feature numbers in Table 4.8.

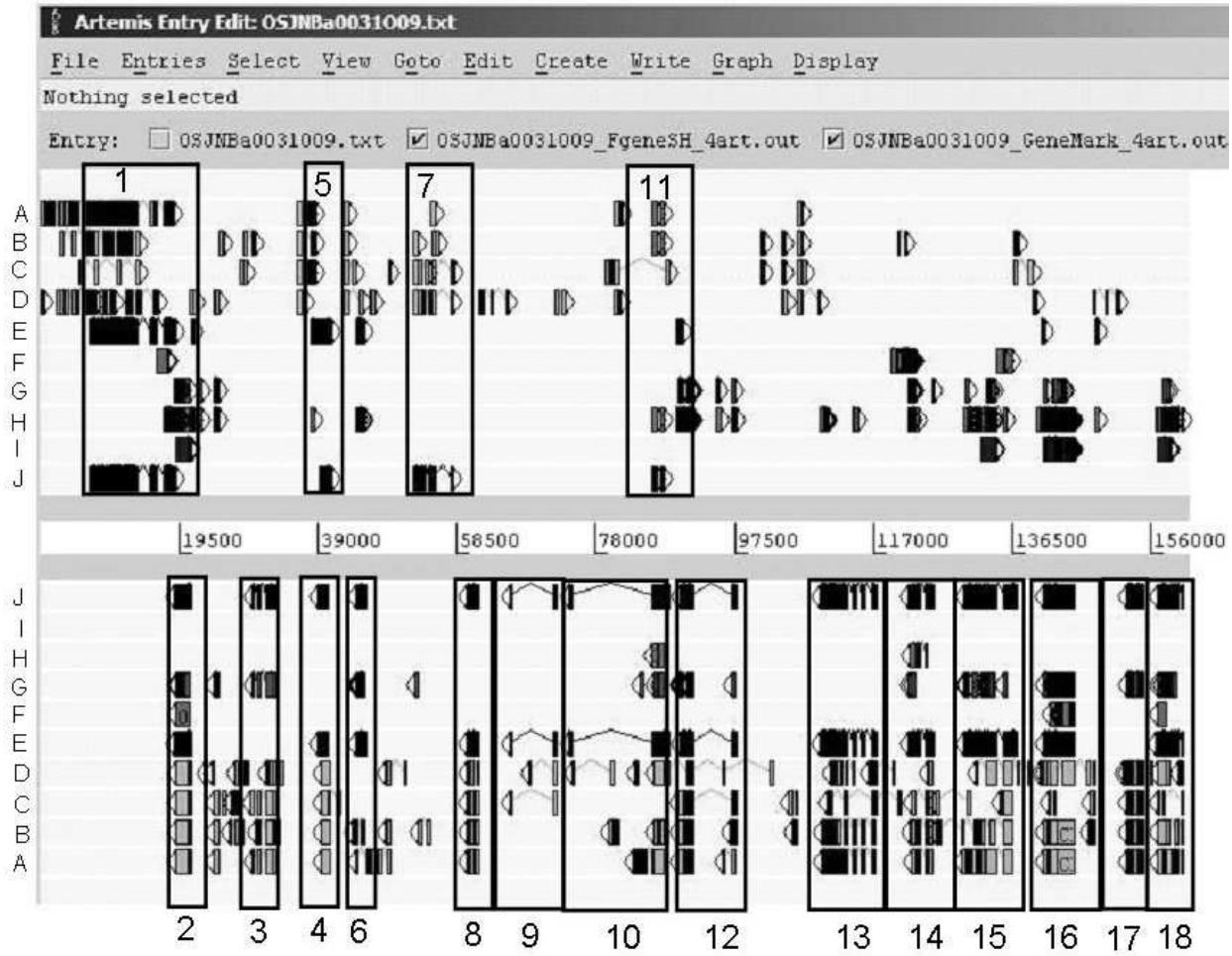


Figure 4.5: Artemis Display of Annotation Evidence on OSJNBa0031O09. The evidence displayed on each track is: gene predictions from (A) FGeneSH, (B) GeneMark, (C) Genscan, (D) RiceHMM. (E) All results generated from a BLASTN comparison of the full-length rice cDNA sequences to the entire OSJNBa0031O09 sequence. Sequence comparisons yielding results with scores greater than 400 and E-values less than 1e-100 from (F) all ORFs greater than 50 amino acids in length compared to the PIR_NREF database, (G) ORFs greater than 50 amino acids in length compared to the non-redundant EST database at NCBI, (H) the BAC sequence broken into overlapping 3 kb sequences compared to the non-redundant EST database, and (I) the BAC sequence broken into overlapping 3 kb sequences compared to the PIR_NREF database. The final assigned annotations are shown on track (J). The numbers of each annotation (1-18) correspond to those feature numbers described in Table 4.6.

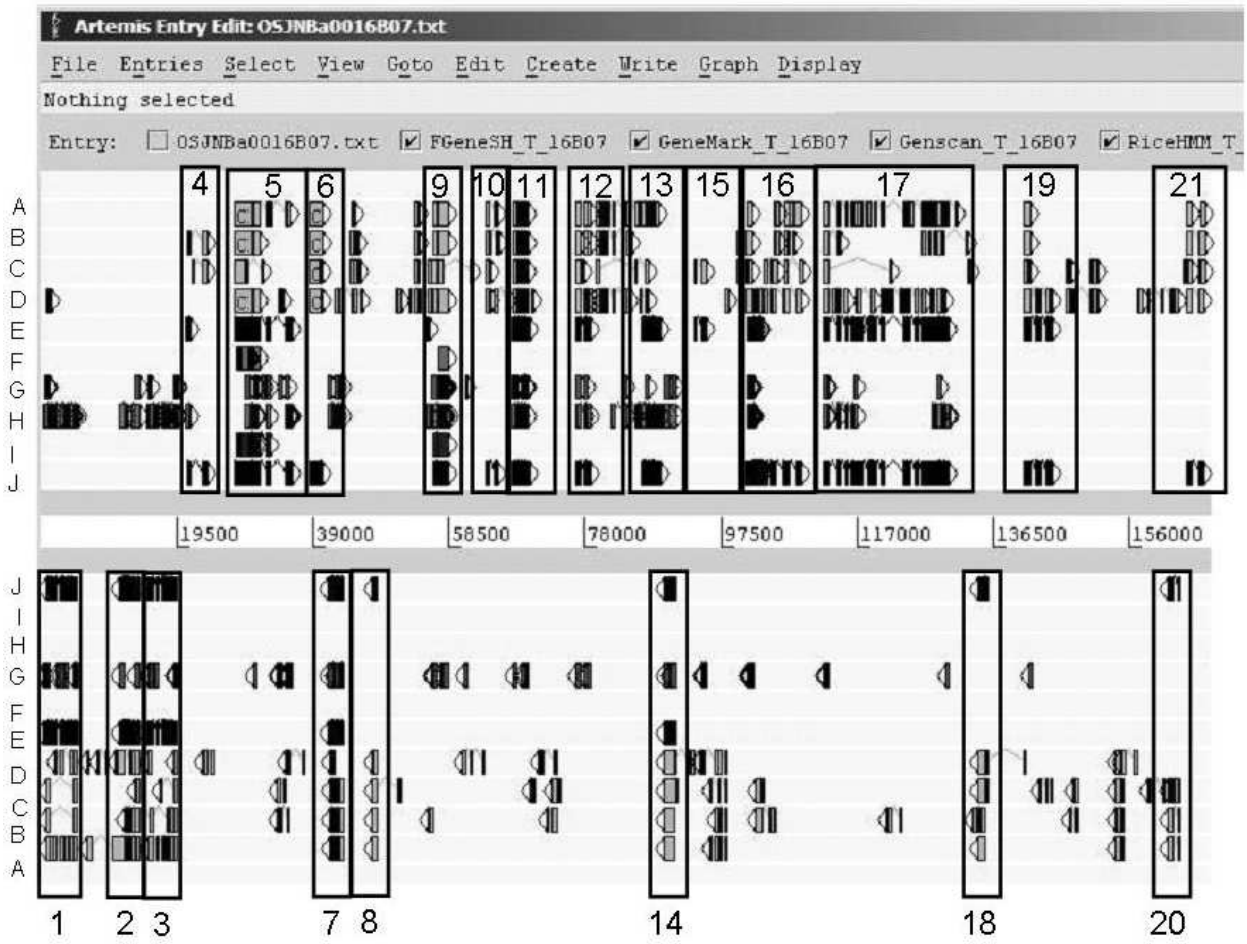


Figure 4.6: Artemis Display of Annotation Evidence on OSJNBa0016B07. The evidence displayed on each track is: gene predictions from (A) FGeneSH, (B) GeneMark, (C) Genscan, (D) RiceHMM. (E) All results generated from a BLASTN comparison of the full-length rice cDNA sequences to the entire OSJNBa0016B07 sequence. Sequence comparisons yielding results with scores greater than 400 and E-values less than 1e-100 from (F) all ORFs greater than 50 amino acids in length compared to the PIR_NREF database, (G) ORFs greater than 50 amino acids in length compared to the non-redundant EST database at NCBI, (H) the BAC sequence broken into overlapping 3 kb sequences compared to the non-redundant EST database, and (I) the BAC sequence broken into overlapping 3 kb sequences compared to the PIR_NREF database. The final assigned annotations are shown on track (J). The numbers of each annotation (1-21) correspond to those feature numbers described in Table 4.7.

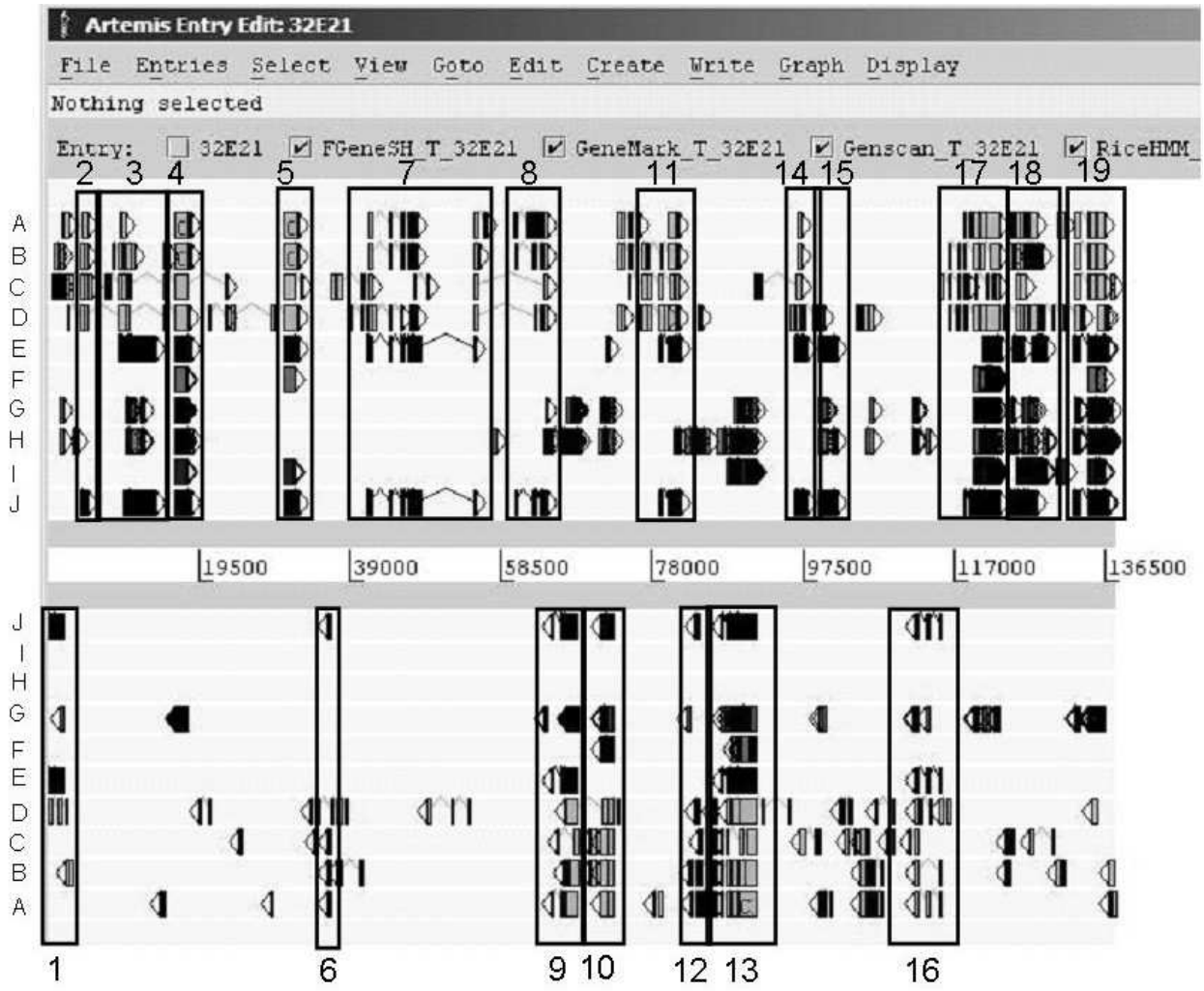


Figure 4.7: Artemis Display of Annotation Evidence on OSJNBa0032E21. The evidence displayed on each track is: gene predictions from (A) FGeneSH, (B) GeneMark, (C) Genscan, (D) RiceHMM. (E) All results generated from a BLASTN comparison of the full-length rice cDNA sequences to the entire OSJNBa0032E21 sequence. Sequence comparisons yielding results with scores greater than 400 and E-values less than 1e-100 from (F) all ORFs greater than 50 amino acids in length compared to the PIR_NREF database, (G) ORFs greater than 50 amino acids in length compared to the non-redundant EST database at NCBI, (H) the BAC sequence broken into overlapping 3 kb sequences compared to the non-redundant EST database, and (I) the BAC sequence broken into overlapping 3 kb sequences compared to the PIR_NREF database. The final assigned annotations are shown on track (J). The numbers of each annotation (1-19) correspond to those feature numbers described in Table 4.8.

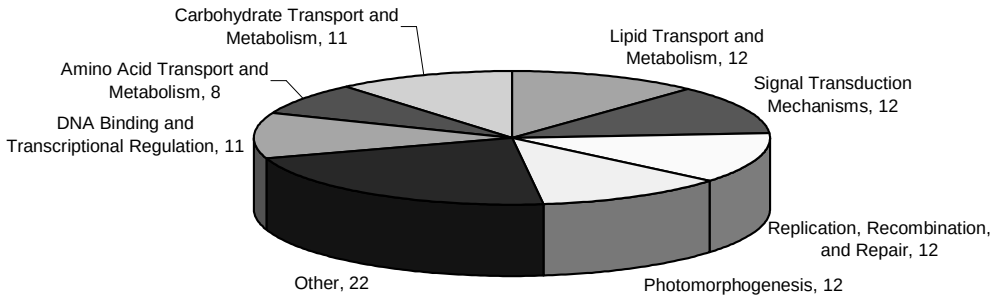


Figure 4.8: Functional Distribution of Identified Genes. The genes identified in the three BAC sequences presented here do not fall into any specific functional classification. The numbers present right of the functional classification indicate the number of genes of that function annotated on the three BACs. The photomorphogenesis genes are only the three *PHY* present in the BAC sequences.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**: 403-410.

Arumunganathan, K., and Earle, E. D. (1991) Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep*., **9**: 208-219.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G. The Gene Ontology Consortium. (2000) Gene ontology: tool for the unification of biology. *Nat Genet.* **25**: 25-9.

Aukerman M. J., Hirschfeld, M., Wester, L., Clack, T., Amasino, R. M., and Sharrock, R. A. (1997) A deletion in the PHYD gene of Arabidopsis Wassilewskija ecotype defines a role for phytochrome D in red/far-red light sensing. *Plant Cell,* **9**: 1317–1326.

Bancroft, I. (2000) Insights into the structural and functional evolution of plant genomes afforded by the nucleotide sequences of chromosomes 2 and 4 of *Arabidopsis thaliana*. *Yeast* **17**: 1-5.

Barbazuk, B., Whitelaw, C., Quackenbush, J., Bennetzen, J., Schubert, K. (2003) The Maize Genome Sequencing Project at the Donald Danforth Plant Science Center. *Maize Genetics Conference Abstracts*. **45**: P16.

Bio-Rad, MicroPulser™ Electroporation Apparatus Operating Instructions and Applications Guide. Biorad Laboratories, Richmond, CA (1994s).

Borodovsky, M. and McIninch, J. (1993) GeneMark: parallel gene recognition for both DNA strands. *Computers & Chemistry* **19:** 123-133.

Borthwick, H.A., Hendricks, S.B., Parker, M.W. (1948) Action spectrum for the photoperiodic control of floral initiation of a long day plant, winter barley (*Hordeum vulgars*). *Bot. Gaz*. **110**: 103-118.

Bray, N., Dubchak, I., Pachter, L. (2003) AVID: A global alignment program. *Genome Res.* **13**: 97-102.

Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol*. **268**: 78-94.

Causse, M.A., Fulton, T.M, Cho, Y.G., Ahn, S.N., Chunwongse, J., Wu, K., Xiao, J., Yu, Z., Ronald, P.C., Harrington, S.E., Second, G., McCouch, S.R., Tanksley, S.D. (1994) Saturated molecular map of the rice genome based on interspecific backcross population. *Genetics* **138**: 1251-1274.

Clack T., Matthews, S., Sharrock, R.A. (1994) The phytochrome apoprotein family in Arabidopsis is encoded by five genes: the sequences and expression of PHYD and PHYE. *Plant Mol. Biol.* **25**: 413-427.

Corbett KD, Berger JM (2003) Emerging roles for plant topoisomerase VI. *Chem Biol.* **10**: 107-111.

Davis, S. J., Kurepa, J., Vierstra, R. D. (1999) The *Arabidopsis thaliana* HY1 locus, required for phytochrome-chromophore biosynthesis, encodes a protein related to heme oxygenases. *Proc Natl Acad Sci, USA,* **96**: 6541–6546.

Dehesh, K., Franci, C., Parks, B. M., Seeley, K. A., Short, T. W., Tepperman, J. M., and Quail, P. H. (1993) Arabidopsis HY8 locus encodes phytochrome A. *Plant Cell,* **5**: 1081–1088.

Devlin, P.F., Robson, P.R.H., Patel, S.R., Goosey, L., Sharrock, R.A., and Whitelam, G.C. (1999). Phytochrome D acts in the shade-avoidance syndrome in Arabidopsis by controlling elongation and flowering time. *Plant Physiol*., **119** ; 909–915.

Donoghue M.J., Mathews, S. (1999) Duplicate genes and the root of angiosperms, with an example using phytochrome sequences. *Mol Phylogenet Evol*. **9**: 489-500.

Egea-Cortines M, Saedler H, Sommer H. (1999) Ternary complex formation between the MADS-box proteins SQUAMOSA, DEFICIENS and GLOBOSA is involved in the control of floral architecture in Antirrhinum majus. *EMBO J*. **18**: 5370-5379.

Gale, M. and Devos, K. (1998) Comparative genetics in the grasses. Proc. Natl. Acad. Sci., USA **95**: 1971-1974.

Honma T, Goto K (2001) Complexes of MADS-box proteins are sufficient to convert leaves into floral organs. *Nature*. **409**: 525-529.

Howe, G.T., Bucciaglia, P.A., Hackett, W.P., Furnier, G.R., Cordonnier-Pratt, M.-M., Gardner, G. (1998) Evidence that the phytochrome gene family in black cottonwood has one *PHYA* locus and two *PHYB* loci but lacks members of the *PHYC/F* and *PHYE* subfamilies. *Mol. Biol. Evol.* **2**: 160-75.

Kolukisaoglu, H. Ü., Braun, B., Martin, W.F., Schnieder-Poetsch, H.A.W. (1993) Mosses do express conventional, distantly B-type related phytochromes: Phytochrome of *Physomitrella patens*. *Hedw. Fed. Eur. Biochem. Soc. Lett*. **334**:95-100.

Lagarias, J. D. (1985) Progress in the molecular analysis of phytochrome. *Photochem. Photobiol*., 42: 811-820.

Lamb RS, Irish VF (2003) Functional divergence within the APETALA3/PISTILLATA floral homeotic gene lineages. *Proc Natl Acad Sci U S A.* **100**: 6558-6563.

Lawrence, C.J, Dong, Q., Polacco, M.L., Seigfried, T.E., Brendel, V. (2004) MaizeGDB, the community database for maize genetics and genomics. *Nucleic Acids Res*., **1**: **Database issue**: D393-397.

Liu LF, Wang JC. (1987) Supercoiling of the DNA template during transcription. *Proc Natl Acad Sci U S A*. **84**: 7024-7027.

Lukashin, A.V. and Borodovsky, M. (1998) GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res*., **26**: 1107-1115.

Lynch, M., O'Hely, M., Walsh, B., Force, A. (2001) The probability of preservation of a newly arisen gene duplicate. *Genetics* **159**: 1789-1804.

Marchler-Bauer A, Anderson JB, DeWeese-Scott C, Fedorova ND, Geer LY, He S, Hurwitz DI, Jackson JD, Jacobs AR, Lanczycki CJ, Liebert CA, Liu C, Madej T, Marchler GH, Mazumder R, Nikolskaya AN, Panchenko AR, Rao BS, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Vasudevan S, Wang Y, Yamashita RA, Yin JJ, Bryant SH (2003) CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res*. **31**: 383-387.

Marchler-Bauer, A., Anderson, J.B., Cherukuri, P.F., DeWeese-Scott, C., Geer, L.Y., Gwadz, M., He, S., Hurwitz, D., Jackson, J.D., Ke, Z., Lanczycki, C.J., Liebert, C.A., Liu, C., Lu, F., Marchler, G.H., Mullokandov, M., Shoemaker, B.A., Simonyan, V., Song, J.S., Thiessen, P.A., Yamashita, R.A., Yin, J.J., Zhang, D., Bryant, S.H. (2005) CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Res*., 1: **Database Issue**: D192-D196.

Marton MJ, Vazquez de Aldana CR, Qiu H, Chakraburtty K, Hinnebusch AG (1997) Evidence that GCN1 and GCN20, translational regulators of GCN4, function on elongating ribosomes in activation of eIF2alpha kinase GCN2. *Mol Cell Biol.* **17***:* 4474-4489.

Mathews, S., Burleigh, J.G., and Donoghue, M.J. (2003) Adaptive Evolution in the Photosensory Domain of Phytochrome A in Early Angiosperms. *Mol. Biol. Evol*., **20**: 1087-1097.

Mathews, S., Lavin, M., Sharrock, R.A. (1995) Evolution of the phytochrome gene family and its utility for phylogenetic analysis of angiosperms. *Ann. Mo. Bot. Gard*. **82**: 266-321.

Mathews, S. and Sharrock, R.A. (1996) The Phytochrome Gene Family in Grasses (Poaceae): A Phylogeny and Evidence that Grasses Have a Subset of the Loci Found in Dicot Angiosperms. *Mol. Biol. Evol*. **13**: 1141-1150.

McCarthy, E.M., Liu, J., Lizhi, G., McDonald, J.F. (2002) Long terminal repeat retrotransposons of *Oryza sativa*. *Genome Biol*. **13**: 3.

Moran TV, Walker JC. (1993(Molecular cloning of two novel protein kinase genes from *Arabidopsis thaliana*. *Biochim Biophys Acta*. **1216**: 9-14.

Morishige DT, Childs KL, Moore LD, Mullet JE. (2002) Targeted analysis of orthologous phytochrome A regions of the sorghum, maize, and rice genomes using comparative gene-island sequencing. *Plant Physiol*. **130**: 1614-25

Muramoto, T., Kohchi, T., Yokota, A., Hwang, I., Goodman, H. M. (1999) The Arabidopsis photomorphogenic mutant hy1 is deficient in phytochrome chromophore biosynthesis as a result of a mutation in a plastid heme oxygenase. *Plant Cell*, **11**: 335–348.

Nagatani, A., Reed, J. W. and Chory, J. (1993) Isolation and initial characterization of Arabidopsis mutants that are deficient in phytochrome A. *Plant Physiol.*, **102**: 269–277.

Ng M, Yanofsky MF (2001) Function and evolution of the plant MADS-box gene family. *Nat Rev Genet*. **2**:186-195.

Parks, B. M. and P. H. Quail (1993) hy8, a new class of Arabidopsis long hypocotyl mutants deficient in functional phytochrome A. *Plant Cell*, **5**: 39–48.

Pasentsis, K., Paulo, N., Algarra, P., Dittrich, P., Thümmler, F. (1998) Characterization and expression of the phytochrome gene family in the moss *Ceratodon purpureus*. *Plant Journal* **13**: 51-61.

Pratt, L.H. (1982) Phytochrome: The protein moiety. *Annu. Rev. Plant Physiol*., **33**: 557-582.

Pratt, L. H., Cordonnier-Pratt, M.-M., Kelmenson, P. M., Lazarova, G., I., Kubota, T., and Alba, R.M. (1997) The phytochrome gene family in tomato. *Plant Cell Environ*. **20**: 672-677.

Quail, P.H. (1991) Phytochrome: A light-activated molecular switch that regulates plant gene expression. *Annu. Rev. Genet*. **25**: 389-409.

Raes, J., Vandepoele, K., Simillion, C., Saeys, Y., Van de Peer, Y. (2003) Investigating ancient duplication events in the *Arabidopsis* genome. *J. Struc. Funct. Gen*. **3**: 117-129.

Reed, J. W., Nagatani, A., Elish, T.D., Fagan, M., Chory, J. (1993) Phytochrome A and phytochrome B have overlapping but distinct functional in Arabidopsis development. *Plant Physiol*., **104**: 1139-1149.

Riechmann JL, Meyerowitz EM (1997). MADS domain proteins in plant development. *Biol Chem*. **378**: 1079-1101.

Roe, B.A., Crabtree, J.S., Khan, A.S. (1996) DNA Isolation and Sequencing. Essential Techniques Series. John Wiley & Sons, New York.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., Barrell, B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics* **10**: 944-945.

Sakata, K., Nagasaki, H., Idonuma, A., Waki, K., Kise, M., and Sasak, T. (1999) A computer program for prediction of gene domain on rice genome sequence. *The 2nd Georgia Tech International Conference on Bioinformatics, Abstracts,* p.78.

Salamov, A. and Solovyev, V. (2000) *Ab initio* gene finding in Drosophila genomic DNA. *Genome Res*., **10**: 516-522.

Salse, J., Piegu, B., Cooke, R., Delseny, M. (2002) Synteny between *Arabidopsis thaliana* and rice at the genome level: a tool to identify conservation in the ongoing rice genome sequencing project. *Nucleic Acids Res*. **30**: 2316-2328.

Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual*., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Sanderson, M.J., and Doyle, J.A. (2001) Sources of error and confidence intervals in estimating the age of angiosperms from rbcL and 18S rDNA data. *Am J. Bot*., **88**: 1499-1516.

Schoof, H. and Karlowski, W. (2003) Comparison of rice and *Arabidopsis* annotation. *Curr. Opin*. *Plant Biol.* **6**: 106-112.

Shah, S.P., McVicker, G.P., Mackworth, A.K., Rogic, S., Ouellette, B.F. (2003) GeneComber: combining outputs of gene prediction programs for improved results. *Bioinformatics* **19**: 1296-1297.

Sharrock, R. A., and Quail, P.H. (1989) Novel phytochrome sequences in *Arabidopsis thaliana*: structure, evolution, and differential expression of a plant regulatory photoreceptor family. *Genes Dev.* **3**: 1745-1757.

Shevchenko, Y., Bouffard, G.G., Butterfield, Y.S.N., Blakesley, R.W., Hartley, J.L., Young, AlC., Marra M. A., Jones, S.J.M., Touchman, J.W., Green, E.D. (2002) Systematic sequencing of cDNA clones using the transposon Tn5. *Nucleic Acids Res*. **30**: 2469-2477.

Smit, AFA, Hubley, R & Green, P. *RepeatMasker Open-3.0*. 1996-2004 <http://www.repeatmasker.org>.

Sonnhammer EL, Durbin R. (1994) A workbench for large-scale sequence homology analysis. *Comput Appl Biosci***. 10**: 301-307.

Stacey G, Koh S, Granger C, Becker JM (2002) Peptide transport in plants. *Trends Plant Sci.* **7**: 257-263.

Tatusova, T.A., Madden, T.L. (1999) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol Lett.* **174**: 247-250.

The Arabidopsis Genome Initiative (2001) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*. 15: 299.

Wang D, Harper JF, Gribskov M. (2003) Systematic trans-genomic comparison of protein kinases between *Arabidopsis* and *Saccharomyces cerevisiae*. *Plant Physiol*. **132**: 2152-2165.

Ware, D., Jaiswal, P., Ni, J., Pan, X., Chang, K., Clark, K., Teytelman, L., Schmidt, S., Zhao, W., Cartinhour, S. (2002) Gramene: A resource for comparative grass genomics. *Nucleic Acid Res.* **30**: 103-105.

Wek RC, Ramirez M, Jackson BM, Hinnebusch AG. (1990) Identification of positive-acting domains in GCN2 protein kinase required for translational activation of GCN4 expression. *Mol Cell Biol*. **10**:2820-31.

Whitelam, G. C., Johnson, E., Peng, J., Carol, P., Anderson, M. L., Cowl, J. S., Harberd, N. P. (1993) Phytochrome A null mutants of Arabidopsis display a wild-type phenotype in white light. *Plant Cell* **5**: 757–768.

Wilson, W.A., Harrington, S.E., Woodman, W.L., Lee, M., Sorrells, M.E., McCouch, S.R. (1999) *Genetics* **153**: 453-473.

Wolfe, K., Guoy, M., Yang, Y., Sharp, P., Li, W.H. (1989) Date of monocot/dicot divergence estimated from chloroplast DNA sequence data. Proc. Natl. Acad. Sci., USA **86**: 5201-5202.

Yeh, K.-C., Wu, S.-H., Murphy, J.T., Lagarias, J.C. (1997) A cyanobacterial phytochrome two-component light sensory system. *Science* **277**: 1505-1508.

Yeh, K.C., Lagarias, J.C. (1998) Eukaryotic phytochromes: light-regulated serine/threonine protein kinases with histidine kinase ancestry.
Proc Natl Acad Sci., U. S. A. **95**: 13976-13981.

CHAPTER 5

INVESTIGATING CONSERVATION AND REGULATORY MOTIFS AMONG *PHY* PROMOTERS

Introduction

The analysis and characterization of the amount of sequence conservation present between two DNA sequences is often used to draw conclusions regarding the expression of genes and the functions of proteins these genes encode. The presence of nucleotide conservation between the coding regions of orthologous genes, defined here as homologous genes in distinct species that have retained similar functions (Mindell and Meyer, 2001), is expected when performing comparative sequence analyses. Such sequences are typically thought to be protected from evolutionary divergence by the functional constraints of protein structure. In contrast to orthologs, genes created through gene duplication within the same genome (Mindell and Meyer, 2001) will not necessarily exhibit such conservation (Rombauts *et al*., 2003). Such genes, known as paralogs, will typically only be retained in the genome if they have acquired novel or complementary functions after duplication (Force *et al*., 1999). Additionally, much literature has been devoted to the existence of sequence conservation outside of coding sequence in non-coding regions of genes (Kaplinsky *et al*., 2002; Guo and Moose, 2003; Choffnes Inada *et al*., 2003; Rombauts *et al.*, 2003). "Phylogenetic footprinting" is the comparative analyses of conserved non-coding DNA sequences (CNSs; Tagle *et al*., 1988; Koop, 1995; Gumucia *et al*., 1996; Duret and Bucher, 1997; and Wasserman *et al*., 2002). CNSs that are sufficiently long to be considered non-random occurrences are hypothesized to contain regulatory sequences known to be involved in the regulation of gene expression.

The poaceae, or angiosperm grass family, includes rice, sorghum, barley, and maize among its more than 10,000 species. Each of these organisms derives from an ancestral genome that existed approximately 50 Mya[1] (Kellogg, 2001). There have only been a handful of studies

[1] Mya=million years ago

concerning the presence of CNSs in the grasses (Kaplinsky *et al*., 2002; Guo and Moose, 2003; Choffnes Inada *et al*., 2003; Rombauts *et al.*, 2003). These studies have typically defined a grass CNS as being either a minimum 15- or 20-nucleotide stretch of DNA sequence conserved with 100% identity.

There are currently multiple algorithms and programs available for the detection of CNSs in both plants and mammals. Various studies have sought to evaluate and refine these tools to determine the best method for detecting CNSs in the grasses (Kaplinsky *et al*., 2002; Guo and Moose, 2003; Choffnes Inada *et al*., 2003; Rombauts *et al.*, 2003). Taken together, these studies offer an overview of the current trend in defining CNSs between orthologous gene pairs in the poaceae.

For example, Kaplinsky *et al*. (2002) defined CNSs between rice and maize as a 15-nucleotide stretch with 100% sequence identity. They determined that the probability of a random 15-nucleotide DNA sequence being passed down, without selection, over 50 million years of grass evolution at a fixed location within the genome was only 4×10^{-6}. Their study utilized the BL2SEQ local alignment algorithm (Tatusova and Madden, 1999) from the National Center for Biotechnology Information (NCBI) to detect CNSs within the genomic sequences of rice and maize (Kaplinsky *et al*., 2002). A second study by Choffnes Inada *et al*. (2003) also used BL2SEQ as described by Kaplinsky *et al.* In addition, however, Choffnes Inada *et al*. developed a Java applet to read the sequence annotation information, align the orthologous sequences being compared, and produce an interactive display of the gene comparisons and CNSs. They, too, defined a CNS as having 15 nucleotides of 100% sequence identity.

In contrast to Choffnes Inada *et al*. and Kaplinsky *et al*., Guo and Moose (2003) determined that 20 nucleotides was the minimal length to identify a significant CNS match

between orthologous genes. Guo and Moose evaluated five tools developed for the detection of sequence similarity in general and/or CNSs specifically (*AVID*: Dubchak *et al*., 2000; Bayes Block Aligner: Zhu *et al*., 1998; DIALIGN: Morgenstern, 1999; DNA Block Aligner: http://www.ebi.acuk/Wise2/dbaforml.html; and PipMaker: Schwartz *et al*., 2000). They determined that of the five tools *AVID* was the most informative in detecting CNS alignments between orthologous gene pairs within the grasses. This study by Guo and Moose also determined that occurrence frequencies for known regulatory elements in CNSs was not elevated above that of random non-conserved sequences within maize-rice orthologous promoters. This finding was in contrast to previously conducted phylogenetic footprinting studies that concluded that CNSs between orthologs are typically enriched in regulatory sequence elements (Levy *et al*., 2001).

Regulatory elements are typically short (*e.g.*, 4-8 nt) sequences and have traditionally been identified through biochemical approaches (Ficket and Hatzigeorgiou, 1997; Bucher, 1999; Sumiyama, Kim, and Ruddle, 2001). They are usually found in the 5' promoter. These sequences are those to which transcription factors bind to orchestrate the selective initiation of transcription (Rombauts *et al*., 2003). The identification of regulatory motifs within CNSs and/or gene promoter sequences, however, is not self-evident. Due to the short nature of such motifs, simply scanning a DNA sequence for their presence will not unambiguously identify them because such short sequences are expected to occur at random every few hundred to tens of thousand base pairs. It is therefore difficult to distinguish between "true" and "false" motifs within a nucleotide sequence (Blanchette and Sinha, 2001). Moreover, it is difficult to evaluate false positive and negative rates when identifying regulatory motifs within conserved sequences because such a small number of regulatory motifs have been experimentally verified. Many

databases containing known regulatory motifs, both general and plant-specific, currently exist. These include EPD (Praz *et al*., 2002), TRANSFAC (Wingender *et al*., 1996), PLACE (Higo *et al*., 1999), and PlantCARE (Lescot *et al*., 2002).

Overview of Study

The objectives of this study are threefold.

Phytochromes are light-responsive proteins encoded by a gene family found in all plants and in the cyanobacterial progenitor of their plastids (Herdman *et al.*, 2000). This study seeks to (a) compare the putative promoter sequences of the rice phytochrome genes to the promoters of their sorghum orthologs to characterize the level of conservation between the promoters of these orthologous gene pairs, and (b) describe the level of conservation, if any, between the upstream promoter regions of three phytochrome paralogs in the rice and sorghum genomes. Regions of conservation in the paralogous and orthologous phytochromes were detected by combining the approaches of Choffnes Inada *et al*. (2003) and Kaplinsky *et al*. (2002) with that of Guo and Moose (2003). Both BL2SEQ and *AVID* were utilized to detect regions of conservation for from 800-2000 bp of 5' promoter sequence of the phytochrome gene family. Moreover, (c) putative regulatory motifs were identified within the rice *PHY*[2] promoters by comparing to databases of known plant promoters those regions of conservation detected between the sorghum and phytochrome orthologs. Entire putative promoter sequences of the three rice *PHY*, whether or not conserved among orthologous *PHY*, were also analyzed for the presence of light regulatory elements (LREs).

[2] *PHY* = phytochrome gene, PHY = phytochrome protein. A specific *PHY* or PHY is designated by the letter(s) of that *PHY*/PHY or *PHY*/PHY subfamily.

Methods

Upstream Promoter Sequences from Rice and Sorghum PHY

Four publicly available *PHY*-containing bacterial artificial chromosome (BAC) sequences for sorghum and rice were downloaded from GenBank (Table 5.1). The location of the *PHY* within each BAC was determined by identifying the start codon of each *PHY* on the BAC sequence by BLASTN[3] (Altschul *et al*., 1990) comparison to complete codon sequences downloaded from GenBank for known rice and sorghum *PHY* (Table 5.1). The SeqEd sequence editor included in the Wisconsin GCG Package version 10.2 (Staden, 1986) was employed to extract at least 2200 nucleotides of sequence upstream of the annotated 5' ATG start codon for the rice *PHYA*-, *B*-, and *C*-containing BACs and 2000 nucleotides upstream of *PHYA* in the sorghum *PHYA*-containing BAC (Table 5.2).

The complete codons of sorghum *PHYB* and *PHYC* (Table 5.1) were downloaded from GenBank and combined with 818 and 1999 nucleotides, respectively, of 5' sorghum *PHYB* and *C* promoter sequences kindly provided by Marie-Michèle Cordonnier-Pratt (unpublished; Tables 5.1 and 5.2).

Eight other genes were downloaded from GenBank (Table 5.1) for use as controls in the comparative analysis of promoter sequences. The two requirements necessary for selection as a control gene were that they not be orthologs or paralogs of *PHY* and must contain at least 1000 nucleotides of annotated 5' upstream sequence (Tables 5.1 and 5.2).

[3] BLASTN parameters: reward for match = 1, penalty for mismatch = 2, cost to open a gap = 5, cost to extend the gap = 2, word size = 11.

Detection of Regions of Sequence Conservation within Putative Promoter Sequences

Three types of comparative analyses were conducted using the putative promoters of the genes described in tables 5.1 and 5.2: paralogous, orthologous, and control (Table 5.3). The paralogous comparisons consisted of comparing the putative rice *PHYA* promoter to the putative promoters of rice *PHYB* and rice *PHYC* and the putative rice *PHYB* promoter to that of rice *PHYC*. This analysis was repeated for the sorghum paralogs *PHYA*, *B*, and *C*. The orthologous comparisons were of the putative promoters of rice *PHYA* to sorghum *PHYA*, rice *PHYB* to sorghum *PHYB*, and rice *PHYC* to sorghum *PHYC*. The three putative promoters from the rice *PHY* were also compared to all eight control genes. These three types of comparative analysis, paralogous, orthologous, and control, result in a total of 33 gene pairs for analysis (Table 5.3).

The 33 gene pairs were compared with both BL2SEQ and *AVID*, yielding a total of 66 comparisons (Table 5.3). The parameters used for BL2SEQ were word size = 7, gap existence penalty = 2, gap extension penalty = 1. Parameters used for *AVID* were window length = 10 nucleotides and conservation level = 90%.

The 66 comparisons detected hundreds of regions of conserved sequences for all three types of gene pairs analyzed. The results were evaluated based on nucleotide length and on the percentage identity of the sequences conserved between two putative promoters.

Each result was designated as one of two types of conserved sequence. (a) Short CNSs (*S*CNSs) were defined as conserved sequences that were at least 10 nucleotides in length with 100% sequence identity or 11-15 nucleotides with at least 90% sequence identity. (b) CNSs were defined as described by Kaplinsky *et al*. (2002) and Choffnes Inada *et al*. (2003). That is, they were defined as conserved regions of 15 nucleotides with 100% sequence identity or of greater than 16 nucleotides with at least 90% sequence identity.

The Artemis Comparison Tool (ACT; http://www.sanger.ac.uk/software/ACT) was used to visualize results of the comparisons (Figure 5.1). MSPCrunch (Sonnhammer and Durbin, 1994) was used to reformat the BL2SEQ results to a format compatible with ACT. *AVID* results were evaluated for regions detected in duplicate. Any regions detected by *AVID* that were not detected by BL2SEQ were added manually to ACT for visualization.

Detecting Regulatory Motifs in the PHY *Upstream Sequences*

Regulatory motifs were detected by utilizing the regulatory motif databases PLACE (plant cis-acting regulatory DNA elements; Higo *et al*., 1999), PlantCARE (Lescot *et al*., 2002), and RegSite (Softberry, Inc., Mount Kisco, NY) through the web interfaces accompanying each database[4].

The analysis of the rice *PHY* promoters through the comparisons of sequences to these databases was performed in four distinct phases from which the final tally of motifs of interest was accumulated.

(1) CNS sequences identified from comparisons of orthologous *PHY* were analyzed. Any regulatory motif identified here was included as a final result.

(2) *S*CNSs conserved in order and orientation between *PHY* detected between orthologous sorghum/rice *PHY* were analyzed for regulatory motifs. Conserved sequences of less than 15 nucleotides, the threshold by which *S*CNSs are defined, are present randomly throughout all genomes. Restraining the motif query to those regions conserved in order and orientation decrease the chance of searching a false positive for motif presence (Bray *et al*.,

[4] The PLACE interface can be found at http://www.dna.affrc.go.jp/PLACE/ and the PLANTCARE interface can be found at http://intra.psb.ugent.be:8080/PlantCARE/. The RegSite web interface can be found at http://www.softberry.com/berry.phtml?topic=regsite.

2003), although this restraint does not ensure that *S*CNSs detected and conserved in order and orientation between two genes is not a false positive. Again, any regulatory motif identified here was included as a final result.

(3) Both the rice and the sorghum sequences that yielded a CNS or *S*CNS not conserved with 100% sequence identity were compared so that the influence of even single-base changes on the presence of regulatory motifs could be detected. For example, a CNS might be detected that is 30 nucleotides long and present at 90% sequence identity between orthologous sorghum and rice promoters. This situation would result in 27 of 30 of the nucleotides being identical. In this case, the 30-nucleotide stretch of the putative promoter sequence from both the sorghum and rice genes would be submitted to the databases for regulatory motif identification in case the 3-nucleotide difference resulted in varying motif detection results.

(4) Putative promoter sequence flanking the CNSs and *S*CNSs detected in rice *PHY* were compared to the same three plant regulatory motif databases to identify any LREs putatively present in the promoters. This search was restricted to the identification of LREs only to decrease the detection of false positive motifs.

Again, the short nature of regulatory motifs makes it difficult to verify their presence within a promoter sequence with any statistical certainty. Biochemically verifying each of the motifs identified here was outside the scope of this project. Therefore, the motifs identified were investigated by a literature review. The focus of this review was on their possible relationship to phytochrome-mediated photoregulation.

Results

Putative promoters of rice and sorghum *PHYA*, *B*, and *C* were compared in pairwise fashion to survey the general characteristics of conserved regions among *PHY* orthologs and paralogs (Table 5.3). Rice *PHY* promoters were also compared to putative promoter sequences described in Tables 5.1 and 5.2 to determine the amount of non-coding sequence similarity between *PHY* and randomly selected control genes (Table 5.3). Thirty-three gene pairs were analyzed for sequence conservation with both BL2SEQ and *AVID*, for a total of 66 comparisons (Table 5.3). The regions of sequence identity detected by BL2SEQ and *AVID* were designated as one of two types, either CNSs or *S*CNSs. The results from both programs were combined. Only one copy of any duplicate entry (those conserved regions detected by both programs) was retained. Conserved sequences and rice *PHY* promoter sequences were then compared to databases of known plant regulatory motifs.

BL2SEQ Detected Substantially More CNSs and SCNSs than AVID

Fourteen CNSs were detected with BL2SEQ, while only seven were detected with *AVID* (Table 5.4). Of the seven detected by *AVID*, three were also detected by BL2SEQ while four were unique to *AVID*. In total, 18 different CNSs were detected in the 66 comparisons. The average length of these CNSs was 21 nucleotides for either program (Table 5.4). BL2SEQ also detected significantly more *S*CNSs than *AVID* (498 versus 201). Again, the average length of the *S*CNSs detected by the two algorithms was identical at 11 nucleotides (Table 5.4).

CNS Occurrence Between Orthologous/Paralogous PHY *is not Higher than in Control Comparisons*

Of the 18 unique CNSs identified by BL2SEQ and *AVID*, only four were between *PHY* orthologs or paralogs (Table 5.5). Specifically, two CNSs were detected between sorghum and rice *PHYC* orthologs and two were detected between the *PHYB* and *PHYC* paralogs of rice.

Within the control comparisons, however, four CNSs were detected between the putative promoters of rice *PHYC* and a NADH II malate dehydrogenase from sorghum, while three CNSs were detected between the rice *PHYC* promoters and a rice transcription factor. In addition, one CNS was detected in each of the *PHYA*/GT1 and *PHYA*/RFL comparisons. In short, CNSs were detected between *PHY* orthologs and paralogs, but not at a significantly greater frequency than between *PHY* and some of the control genes (*e.g.*, Table 5.5, compare gene pairs 9 and 24). Previous studies have revealed that many orthologous gene pairs between rice and maize (27% of the 52 studied) did not reveal any CNSs when a CNS was defined as 15-nucleotide match of 100% identify (Kaplinsky *et al*., 2002). Thus, the lack of conserved sequences longer than 15 nucleotides between the rice/sorghum *PHY* orthologs is not necessarily an anomaly.

In contrast, *PHY* orthologs did display a higher frequency of *S*CNSs than any other type of gene pair analyzed (Figure 5.2 and Table 5.5). The numbers of *S*CNSs detected per type of gene pair were normalized with respect to the amount of promoter sequence analyzed for each type of gene pair (Figure 5.2). Typically, if a CNS was detected then a higher number of *S*CNSs could be expected suggesting the overall conservation of gene pairs with detected CNSs is higher than in non CNS-containing gene pairs (Table 5.1). These detected *S*CNSs, however, are not long enough in nucleotide length to be declared individually as present with any statistical significance.

Light Regulatory and Auxin-Response Elements are the Most Frequent in the Conserved Regions of the Putative PHY *Promoters*

Regulatory motifs were identified for othologous *PHY* within both CNSs and *S*CNSs by comparison to three databases of known plant promoters (Figures 5.3, 5.4, and 5.5).

Twenty-four sequences, representing the CNSs and *S*CNSs conserved within putative 5' promoter region of orthologous rice and sorghum *PHY*, were scanned for the presence of known regulatory motifs. Within these conserved regions, 37 regulatory motifs were identified (Figure 5.6). It is not known whether any of these are functional as they have not been experimentally verified. The descriptions of all motifs detected are listed in Table 5.6.

Of those 37 motifs identified within the CNSs or *S*CNSs of the rice *PHY*, 10 were known LREs. These LREs were the most abundant class of motifs identified within the CNSs and *S*CNSs (Figure 5.6). In addition, 27 motifs of other function were found (Figure 5.6). For example, all three *PHY* promoters of rice also contain a circadian rhythm LELHC motif (PLACE database ID: Circadian LELHC, S000252; Figures 5.3, 5.4 and 5.5). LELHC is a consensus sequence necessary for the circadian expression of the tomato light harvesting complex protein (Lhc) gene (Piechulla *et al*., 1998). Although this motif is present in all three rice *PHY* promoters, the motif is located at varying distances upstream from the +1 coding ATG in the different *PHY* (Figures 5.3, 5.4, and 5.5).

The most predominant regulatory motifs detected in the conserved regions of the *PHY* promoters, other than LREs, were those involved in auxin response. Nineteen percent of all regulatory motifs detected were identified as being involved in auxin response. One of these motifs, PLACE database ID ASF1 Motif CAMV, S000024, is located in the same relative position approximately 500 to 550 nucleotides upstream of the ATG +1 start site in the *PHYA*

and *PHYC* promoter regions (Figures 5.4 and 5.6, respectively), while it is located more downstream at approximate position -250 in the *PHYB* promoter (Figure 5.5).

In addition to the ten LREs identified within the CNSs and *S*CNSs of rice and sorghum *PHY* orthologs, when the entire rice *PHY* promoters were compared to the plant motif databases 16 other LREs were detected in regions outlying the conserved sequences (Figures 5.3, 5.4, and 5.5). The LREs present in the *PHY* promoters are mainly I boxes (PLACE database ID: I box, S000124), GATA Boxes (PLACE database ID: GATA, S000039), and GT1 Consensus Sequences (PLACE database ID: GT1, S000198). Due to the short nature of these motifs it cannot be determined whether their presence is significant.

Discussion

The Number of Conserved Regions Detected is a Function of the Alignment Method

Approximately twice as many CNSs and *S*CNSs were detected by BL2SEQ as compared to *AVID*, although the average length of both types of conserved sequences was identical regardless of which algorithm was employed (Table 5.4). Specifically, 14 CNSs were detected by BL2SEQ while 7 were detected by *AVID* (Table 5.4). Three of the 7 CNSs detected by *AVID* were also detected by BL2SEQ.

The increased number of conserved regions detected by BL2SEQ over *AVID* can be explained by differences between these two algorithms. *AVID* is a global alignment algorithm that assumes that biologically significant regions of sequence similarity are present in conserved

order and orientation (Bray et al., 2003). In contrast, BL2SEQ is a local alignment algorithm that does not make this assumption (Tatusova and Madden, 1999).

The CNSs detected by *AVID* and BL2SEQ do not appear to be more relevant than those detected by BL2SEQ alone. Regulatory motifs were not present at a higher frequency in these CNSs than those detected solely by BL2SEQ.

The Extent of Conservation is Correlated with Gene Function

The functions of genes in a gene pair played a substantial role in the number of conserved regions observed between promoters. The number of CNSs and *S*CNSs was substantially higher when *PHY* promoters from rice were compared to a promoter from malate dehydrogenase, a light-responsive gene, and *RFL,* a transcription factor known to be involved in *PHY* signaling, than when *PHY* was compared to control genes encoding proteins of other functions.

Specifically, when the promoters of the three rice *PHY* were compared to NADH II malate dehydrogenase, nine distinct CNSs were detected (Table 5.5, Gene Pairs 28-30). Four of these nine were detected between rice *PHYC* and NADH II malate dehydrogenase (Table 5.5, Gene Pair 30). This comparison yielded the highest number of CNSs detected from any gene pair analyzed in this study. Interestingly, the expression of NADH II malate dehydrogenase, like *PHY*, is influenced by environmental light levels (Miginiac-Maslow *et al.*, 1997). This light responsiveness might explain why the level of conservation is high between these promoters as both contain sequences necessary to respond to light stimuli. Also, three CNSs were detected between *PHYC* and *RFL* (Table 5.5, Gene Pair 24), the latter a known plant transcriptional activator (Prasad *et al*., 2003).

In contrast, when the rice *PHY* were compared to genes encoding storage proteins (Table 5.5, Gene Pairs 31-33), structural proteins (Table 5.5, Gene Pairs 25-27), and proteins from evolutionarily distant organisms that exhibit no discernable connection to *PHY* (Table 5.5, Gene Pairs 10-18,) there were no CNSs and very few *S*CNSs detected.

These findings are consistent with those of Choffnes Inada *et al*. (2003), who also made the observation that gene function contributes substantially to the number of CNSs detected between orthologous gene pairs. They, too, found that when genes encoding structural proteins were compared to other grass genes few, if any, detectable CNSs were identified. In contrast, genes encoding transcription factors or proteins hypothesized to be involved in complex regulatory interactions were typically enhanced in CNSs (Choffnes Inada *et al*., 2003).

LREs Are the Most Abundantly Detected Motif Within the Conserved Regions of Putative Rice PHY *Promoters*

Ten LREs were detected within the regions of conservation of putative rice *PHY* promoters. These constitute 27% of all motifs detected in these regions. Moreover, 16 additional LREs were detected in regions flanking the CNSs and *S*CNSs. LREs are known to be essential for light-controlled transcriptional regulation and commonly occur in the promoters of light-regulated genes. Examples of such LREs are the G-, GATA-, GT1-boxes (Millar and Kay, 1996; Terzaghi and Cashmore, 1995; Tobin and Kehoe, 1994), all of which are present in the putative promoters of the rice *PHY* shown here (Figures 5.3, 5.4, and 5.5).

PHYA expression is considered to be strongly photoregulated at the mRNA level (Quail, 1994). A number of studies that describe the steady-state levels of *PHYA* mRNA have been published for Arabidopsis (Clack *et al*., 1994), potato (Heyer and Gatz, 1992 a and b), tobacco, (Ádám *et al*., 1994) and tomato (Hauser *et a*l., 1997 and 1998). These studies all report the

presence of photoregulation of *PHYA* mRNA. The presence of photoregulation in the remaining phytochromes has, however, been more debated. Studies in all of the above plants (Heyer and Gatz, 1992 a and b; Ádám *et al*., 1994; Clack et al., 1994), except tomato, conclude that *PHYB* is not photoregulated at the mRNA level. In contrast, studies measuring quantitatively the levels of *PHY* transcripts in tomato (Hauser et al., 1998) indicate that *PHYB* is photoregulated, just on a lesser than scale than for *PHYA*.

While studies have been performed that described certain regulatory motifs inherent to *PHY* promoters (Bruce *et al*., 1990, Bruce and Quail, 1990, Bruce *et al*., 1991, Dehesh *et al*., 1994, Morishige *et a*l., 2002), none have specifically addressed the presence or absence of LREs in the *PHY* promoters. As stated above, *PHYA* has been shown to be highly photoregulated in a number of plants, so LREs might be expected in this promoter. Does the rice *PHYA* promoter described here contain different LREs than *PHYB* and *PHYC* that might confer this photoregulation? It was demonstrated in this study that *PHYA* and *PHYC* both contained LREs in the regions conserved between the rice/sorghum *PHY* orthologs, while *PHYB* only contained LREs in sequence flanking the detected regions of conservation. *PHYA* was shown to contain two GATA boxes, two GT1 consensus sequences, and one CAAT box within the CNSs and *S*CNSs identified. *PHYC* contained three GATA boxes, one I-box, 1 GT1 consensus sequence, one CAAT box, and one RBCS consensus sequence within the CNSs and *S*CNSs identified (Figures 5.3, 5.4, and 5.5). It has previously be shown using G, GATA, and GT1 LREs that paired element-containing promoters respond to a broad spectra of ambient light, yet promoters containing a single LRE only respond to lights of a specific wavelength (Chattopadhyay *et al*., 1998b). It is possible that photoregulation is conferred via a specific combination of LREs present in *PHYA* promoters, but this supposition cannot be unambiguously confirmed here.

Regardless, the results presented here, namely the presence of multiple LREs within regions conserved in *PHY* rice/sorghum orthologs; suggest LREs might be important to the *PHY*.

Multiple studies have recently investigated the importance of the combinatorial effects LREs (Degenhardt and Tobin, 1996; Feldbrugge *et al*., 1997; Puente *et al*., 1996). The native responses of promoters that respond to phytochrome activating light pulses can be imitated by manufactured promoters containing LREs present in pairs. These promoters also respond to developmental signals regulating responses such as chloroplast development (Puente *et al*., 1996). Paired G-, GATA-, and GT1-containing promoters respond to a broad spectrum of light, while promoters containing only a single LRE sometimes respond to very specific wavelengths of light (Chattopadhyay *et al*., 1998b). Although determining the biochemical responses of the combinations of LREs is outside the scope of this project, the identification of LREs present in putative promoter sequences for directed experimental verifications would facilitate this type of combinatorial analysis.

Auxin Response Elements Are the Second Most Abundant Motif Type in the Putative PHY *Promoters*

Plant cells communicate over long distances by transporting small signaling molecules such as auxin (Jürgens, 1993). It is known that auxin regulates various aspects of plant growth and development such as cell elongation and division, organ and tissue differentiation, and morphogenesis (Jürgens, 1993). These responses are activated and regulated by auxin via signal transduction mediated by activation of a specific group of transcription factors (Abel and Theologis, 1996).

Other than LREs, auxin response elements (AREs) were the second most abundant type of regulatory motif identified here. Why should a *PHY* promoter have elements in common with

an auxin-inducible promoter? Is there a relationship between auxin response and *PHY*, or is it just that auxin promoters are among the best-described promoters?

A number of situations are described in the literature where auxin- and phytochrome-induced responses in plants are linked (for review, see Swarup *et al*., 2002). It is known that several auxin-inducible genes are up-regulated in wild-type Arabidopsis in response to a reduction in light intensity (Vandenbussche *et al*., 2003). Also, a number of auxin-response proteins involved in a transduction chain that leads to an auxin response can be phosphorylated by PHYA (Colon-Carmona *et al*., 2000). Therefore, the light-signaling auxin pathways are clearly intertwined (Vandenbussche *et al*., 2003).

A third example of a phytochrome/auxin relationship is highlighted by the shade avoidance phenomenon seen in plants in response to changes in the ambient light spectrum that is observed within forest canopies (Holmes, 1983). These changes include a decrease in light intensity, in particular of blue and red light, both of which influence photosynthesis and photomorphogenesis. To overcome these changes many plants can adapt their phenotype to 'reach out' for light and avoid shading (Holmes, 1983; Smith and Whitelam, 1997). Some plants redirect the accumulation of biomass to the stem and petiole region instead of leaf blades (Smith, 1992; Hangarter, 1997; Malikal *et al*., 1999). Auxin is one plant hormone that can induce the reorientation of leaf blades (Brock *et al*., 1994; Clua *et al*., 1996; Cox *et al*., 2003) and modify the rate of hypocotyl elongation (Steindler *et al*., 1990). PHYB mutants also have elevated petioles even in non-shaded conditions (Somers *et al*., 1999) and display the same elongated hypocotyls and petioles as seen in wild type plants grown in shaded conditions (Koornneef *et al*., 1980; Reed *et al*., 1993; Devlin *et al*., 1996).

The simple presence of AREs in the *PHY* promoters is not, however, automatically explained by the observation that both auxin and phytochrome can have similar effects. The presence of AREs in *PHY* promoters suggests that *PHY*, in some way, is controlled by auxin. Simply, this control could result in one of two outcomes. Auxin could influence a photomorphogenic response by inducing or repressing *PHY* and the subsequence PHY-mediated response. Or, auxin could impart some *PHY* control which then imparts some PHY-mediated "auxin" response. Although a link between auxin- and *PHY*-mediated responses can be established through the literature, determining whether the AREs described here are simply a function of the highly-characterized nature of auxin response motifs, or are indicative of a verifiable link between photomorphogenic *PHY*-mediated responses and auxin response remains to be seen.

Conclusion

Employing a comparative genomics approach to reveal patterns of noncoding sequence conservation and evolution, particularly in promoter regions, can offer valuable insights into the complex regulation of gene expression in plants. Although identifying the regulatory motifs present within these regions and surrounding promoter sequence is not self-evident, the strategies described here will become an increasingly powerful approach to elucidating mechanisms of plant gene regulation as more plant genes and genomic regions are sequenced and made public. Unfortunately, the lack of both putative and experimentally defined promoter sequences from *PHY* in multiple plant species hinders the generation of an abundance of sequence for meaningful comparisons. Also, the short nature of the sequences conserved among the orthologous and paralogous *PHY* make it very difficult to statistically verify their significance.

The brief length of known regulatory motifs further compounds this problem. Regardless, the identification of short sequences conserved between the promoters of orthologous genes from multiple species will be invaluable in guiding laboratory experiments to verify the action of plant regulatory motifs.

Figures and Tables

Table 5.1: Description of Sequences Downloaded from GenBank for Use in this Study.

Sequence Name*	GenBank Accession	Description	Reference
OSJNBa0031O09	AF377946	*PHYA*-containing BAC from rice	This thesis
OSJNBa0016B07	AF461424	*PHYB*-containing BAC from rice	This thesis
OSJNBa0032E21	AF377947	*PHYC*-containing BAC from rice	This thesis
Sbb3766	AF369906	*PHYA*-containing BAC from *Sorghum bicolor*	Morishige *et al*., 2002
Os PHYA	AB109891	*Oryza sativa PHYA* complete codons	Tahir *et al*., 2003, Published Only in GenBank
Os PHYB	AB109892	*Oryza sativa PHYB* complete codons	Tahir *et al*., 2003, Published Only in GenBank
Os PHYC	AF141942	*Oryza sativa PHYC* complete codons	Basu *et al*., 2000
Sb PHYA	AY466073	*Sorghum bicolor PHYA* complete codons	White *et al*., 2004
Sb PHYB	AF182394	*Sorghum bicolor PHYB* complete codons	Alba *et al*., 2000
Sb PHYC	*AY466458*	*Sorghum bicolor PHYC* complete codons	White *et al*., 2004
Dm hsp70b	AY370940	Drosophila heat shock gene, promoter, and partial codons	none
Rs Pbsn	AY370611	Rat probasin gene, promoter, and complete codons	Kasper and Matusik, 2000
Om Differentiation 6-1 gene	AY325275	Oncorhynchus differentiation promoter and complete codons	none
Os GT1	AY338469	Glutathione trasporter; gene, promoter and partial codons	Unpublished
Os RFL	AF397034	Rice transcription factor promoter and partial codons	Prasad *et al*., 2003
Os BP-73	AJ315790	Promoter and partial codons	Chen *et al*., 2003
Sb NADH malate dehydrogenase	X54404	Sorghum malate dehydrogenase gene and promoter	Unpublished
Sv GK	X62480	Complete codons and promoter of protein storage gene from sorghum	de Freitas *et al*., 1994

**Species are abbreviated as follows:*
Os: Oryza sativa *Sb: Sorghum bicolor* *Om: Oncorhynchus mykiss*
Sv: Sorghum vulgare *Dm: Drosophila melanogaster* *Rn: Rattus norvegicus*

Table 5.2: Function and Promoter Lengths of Genes Used in Comparative Analyses. The table below describes the function of the proteins encoded by the genes used in the comparative analyses and the length of their respective putative promoter sequences.

Gene	Protein Function	Length of putative promoter available for analysis
OsPHYA	Light-regulated photoreceptor	2200
Os PHYB	Light-regulated photoreceptor	2200
Os PHYA	Light-regulated photoreceptor	2200
Sb PHYA	Light-regulated photoreceptor	2000
Sb PHYB	Light-regulated photoreceptor	1246
Sb PHYC	Light-regulated photoreceptor	1999
Dm hsp70b	heat shock protein	1607
Rs Pbsn	Probasin protein in epithelial cells of rat prostate	2000
Om Differentiation 6-1 gene	unknown	1746
Os GT1	Glutathione transporter	2247
Os RFL	Transcription factor in developing rice inflorescence	3516
Os BP-73	DNA-binding protein putatively involved in RNA metabolism	2023
Sb NADH malate dehydrogenase	light-activated chloroplast enzyme	1111
Sv GK	Gamma kafirin: Storage protein	1224

**Species are abbreviated as follows:*

Os: Oryza sativa *Sb: Sorghum bicolor* *Om: Oncorhynchus mykiss*
Sv: Sorghum vulgare *Dm: Drosophila melanogaster* *Rn: Rattus norvegicus*

Table 5.3: Analyzed Gene Pairs. Thirty-three gene pairs were analyzed by two methods each for a total of sixty-six comparisons.

Comparison Type	Sequence 1	Sequence 2	Number of Gene Pairs Analyzed	Number of Analyses Performed
Control	*Os PHY**	hsp70bp	3	6
Control	*Os PHY**	Pbsn	3	6
Control	*Os PHY**	Differentiation 6-1 gene	3	6
Control	*Os PHY**	GT1	3	6
Control	*Os PHY**	*RFL*	3	6
Control	*Os PHY**	BP-73	3	6
Control	*Os PHY**	NADH II malate dehydrogenase	3	6
Control	*Os PHY**	GK	3	6
Orthologous	*Os PHYA*	*Sb PHYA*	1	2
Orthologous	*Os PHYB*	*Sb PHYB*	1	2
Orthologous	*Os PHYC*	*Sb PHYC*	1	2
Paralogous	*Os PHYA*	*Os PHYB*	1	2
Paralogous	*Os PHYB*	*Os PHYC*	1	2
Paralogous	*Os PHYA*	*Os PHYC*	1	2
Paralogous	*Sb PHYA*	*Sb PHYB*	1	2
Paralogous	*Sb PHYB*	*Sb PHYC*	1	2
Paralogous	*Sb PHYA*	*Sb PHYC*	1	2

* All three *PHY* from rice (*PHYA*, *B*, and *C*) were compared to the control genes.

Table 5.4: Summary of Comparison Results. Total number of CNSs and *S*CNSs detected by BL2SEQ and *AVID* and the average lengths of each type of detected region of conservation.

Algorithm	Total Number of CNSs	Average Length of a CNS (nucleotides)	Total Number of *S*CNSs	Average Length of a *S*CNS (nucleotides)
BL2SEQ	14	21	498	11
AVID	7	21	201	11

Table 5.5: Summary of CNSs and *S*CNSs Detected in Each Type of Comparison.

Gene Pair Number	Comparison Type	Sequence 1	Sequence 2	No. CNS detected	No. *S*CNS detected
1	Paralogs	*OsPHYA*	*OsPHYB*	0	12
2	Paralogs	*OsPHYB*	*OsPHYC*	2	18
3	Paralogs	*OsPHYA*	*OsPHYC*	0	12
4	Paralogs	*SbPHYA*	*SbPHYB*	0	14
5	Paralogs	*SbPHYB*	*SbPHYC*	0	18
6	Paralogs	*SbPHYA*	*SbPHYC*	0	10
7	Orthologs	*OsPHYA*	*SbPHYA*	0	24
8	Orthologs	*OsPHYB*	*SbPHYB*	0	24
9	Orthologs	*OsPHYC*	*SbPHYC*	2	33
10	Control-heat shock	*OsPHYA*	hsp70bp	0	12
11	Control-heat shock	*OsPHYB*	hsp70bp	0	13
12	Control-heat shock	*OsPHYC*	hsp70bp	0	11
13	Control-Pbsn	*OsPHYA*	Pbsn	0	14
14	Control-Pbsn	*OsPHYB*	Pbsn	0	10
15	Control-Pbsn	*OsPHYC*	Pbsn	0	12
16	Control-Dif	*OsPHYA*	Differentiation 6-1 gene	0	10
17	Control-Dif	*OsPHYB*	Differentiation 6-1 gene	0	12
18	Control-Dif	*OsPHYC*	Differentiation 6-1 gene	0	14
19	Control-GT1	*OsPHYA*	GT1	1	15
20	Control-GT1	*OsPHYB*	GT1	0	8
21	Control-GT1	*OsPHYC*	GT1	0	16
22	Control-RFL	*OsPHYA*	*RFL*	1	22
23	Control-RFL	*OsPHYB*	*RFL*	0	18
24	Control-RFL	*OsPHYC*	*RFL*	3	15
25	Control-BP73	*OsPHYA*	BP-73	0	9
26	Control-BP73	*OsPHYB*	BP-73	0	13
27	Control-BP73	*OsPHYC*	BP-73	0	15
28	Control-malate	*OsPHYA*	NADH II malate dehydrogenase	3	22
29	Control-malate	*OsPHYB*	NADH II malate dehydrogenase	2	19
30	Control-malate	*OsPHYC*	NADH II malate dehydrogenase	4	24
31	Control-GK	*OsPHYA*	Gamma Kafarin	0	8
32	Control-GK	*OsPHYB*	Gamma Kafarin	0	10
33	Control-GK	*OsPHYC*	Gamma Kafarin	0	12

Table 5.6: Motifs Detected in Rice *PHY* genes. The following table is an alphabetized list of all motifs with their respective sequences, references, and functions of all motifs detected in the phytochrome genes.

Motif Name	Sequence	Reference	Function/Comment
ASF1 Motif	TGACG	Lam *et al*., 1989;	TGACG motifs are found in many promoters and are involved in
CAMV		Katagari *et al*., 1989;	transcriptional activation of several genes by auxin and/or salicylic acid;
		Klinedinst *et al*., 2000	May be relevant to light regulation
CAAT Box	CCAAT	None	CBP binds to CCAAT boxes to stabilize the RNA POLII complex. Found in the upstream regions (~ -80) of many eukaryotic genes
CATATGG MS AUR	CATATG	Xu *et al*., 1997	Involved in auxin responsiveness in soybean SAUR (Small Auxin-Up RNA) 15A gene promoter.
CGACG OS Amy3	CGACG	Hwang *et al*., 1998	Found in the GC-rich regions of rice Amy3D and Amy3E amylase genes.
Circadian LELHC	CAANNNN ATC	Piechulla *et al*., 1998	Sequence necessary for the circadian expression of tomato Lhc gene.
DOF Core	AAAG	Yanagisawa and Schmidt, 1999	DOF binding protein site. DOF binding proteins are known to both enhance and repress transcription
E Box	CANNTG	Stalberg *et al*., 1996	The disruption of an overlapping E box motif abolished the transcription of the napA storage protein in *Brassica nappus.*
GATA Box	GATA	Teakle *et al*., 2002	Known LRE. Binds with ASF-2. Required for high-level light-regulated and tissue-specific expression of chlorophyll a/b finding protein in petunia.
GT1 Consensus	GRWAAW	Villain *et al*., 1996; Le Gourrierec *et al*., 1999; Buchel *et al*., 1999	Consensus GT-1 binding site in many light-regulated genes. The activation of GT-1 may be achieved through the direct interaction of TFIIA and GT-1.
I Box	GATAAG	Rose *et al*., 1999	Conserved sequence upstream of light-regulated genes. Binding site of LeMYB2 which is a novel class of myb-like proteins and a known transcriptional activator
Myb Core	CNGTTR	Urao *et al*., 1993	Binding site for ATMYB2 and ATMYB1 in Arabidopsis. ATMYB2 is known to be involved in regulating genes responsive to water stress.
MYBST1	GGATA	Baranowskij *et al*., 1994	A novel DNA binding protein with homology to Myb oncaproteins containing only one repeat can function as a transcriptional activator The Myb motif of the MybSt1 protein is distinct from the plant Myb DNA binding domain.
NTBBF ARROLB	ACTTTA	Baumann *et al*., 1999	A tobacco DOF binding site in an agrobacterium rolB gene. Required for the tissue-specific expression of auxin induction.
POLA Sig2	AATTAAA	O'Neill *et al*., 1990	Consensus sequence for plant polyadenylation signal
POLA Sig3	AATAAT	Joshi, 1997	Consensus sequence for plant polyadenylation signal
Pollen LeLat 52	AGAAA	Bate and Twell, 1998	One of two co-dependent elements responsible for the pollen specific activation of the tomato lat52 gene.
RAV1AAT	CAACA	Kagaya *et al*., 1997	Binding consensus of Arabidopsis TF RAV1. The expression of RAV1 in Arabidopsis are relatively high in rosette leaves and roots.
RBCS Consensus	AATCCAA	Donald and Cashmore, 1990	Known light-responsive element.
RE Alpha LGLHCB21	AACCAA	Degenhardt and Tobin, 1996	Required for phytochrome regulation in the lemma gibba lgcb21 promoter. The DNA binding activity is high in etiolated plants but much lower in green plants.
RE Beta LGLHCB21	CGGATA	Degenhardt and Tobin, 1996	Required for phytochrome regulation in the lemma gibba Lhcb21 gene promoter
Root Motif	ATATT	Elmayan and Tepfer, 1995	Found in both promoters of rolD which is strongly expressed in roots.
SEF4	RTTTTR	Lessard *et al*., 1991	SEF4 binding site found in soybean beta-conglycinin, a seed storage protein.
TATA Box4	TATATAA	None	A functional TATA element by *in vivo* analysis in sweet potato

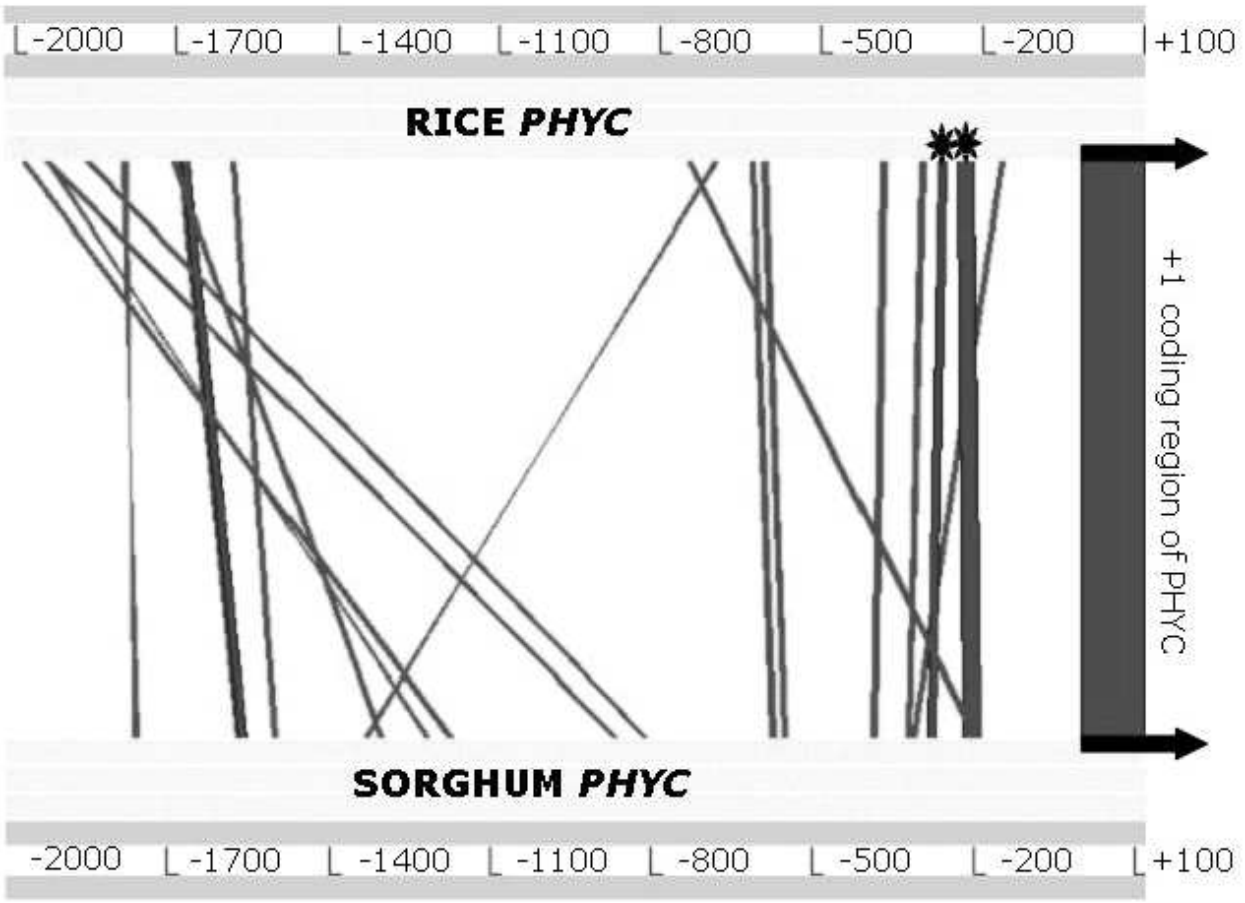


Figure 5.1: ACT View of Regions of Sequence Similarity Detected Between the *PHYC* Promoters of Rice and Sorghum by Both BL2SEQ and *AVID*. Stars (✷) identify CNSs. The rest of the regions displayed here are *S*CNSs. The numbers across the top and bottom of the figure represent the numbering of the promoter in nucleotides with the ATG start site being designated as +1.

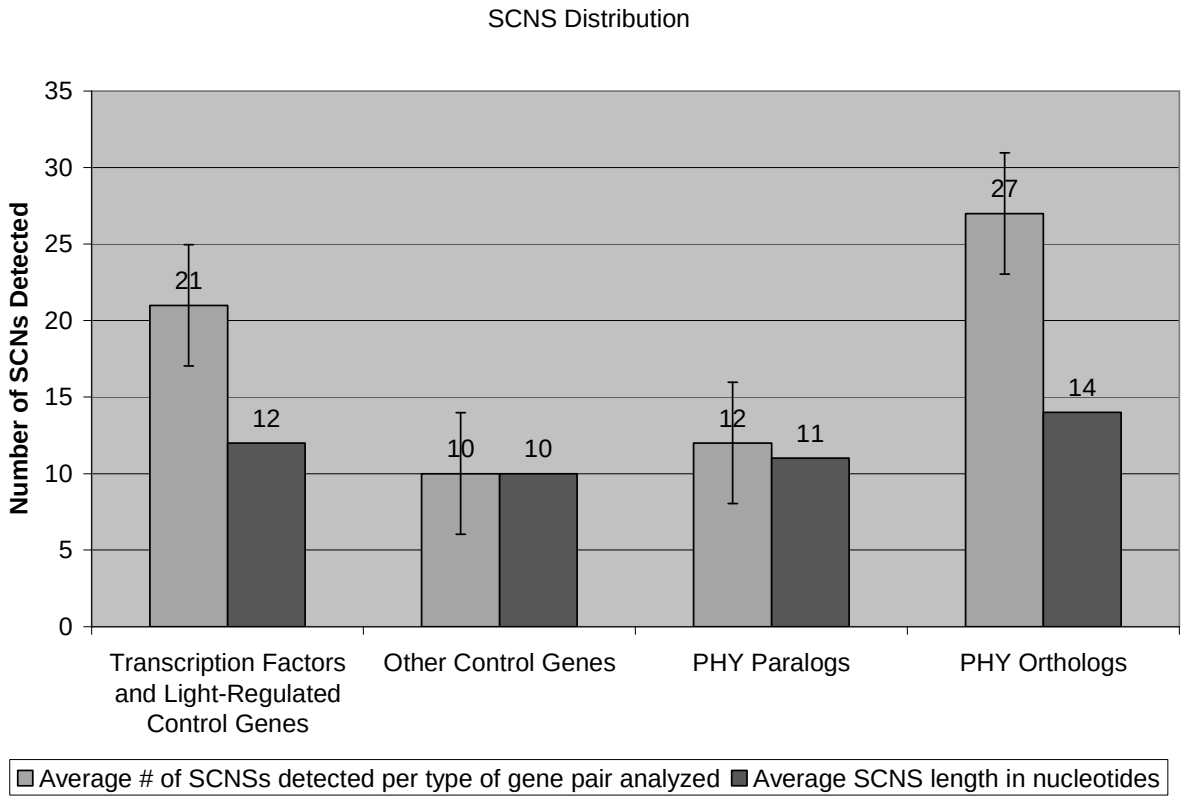


Figure 5.2: Average Number and Length of *S*CNSs Detected Per Type of Gene Pair Analyzed. The mean *S*CNSs detected and reported here have been normalized with respect to the amount of promoter sequence analyzed per type of gene pair.

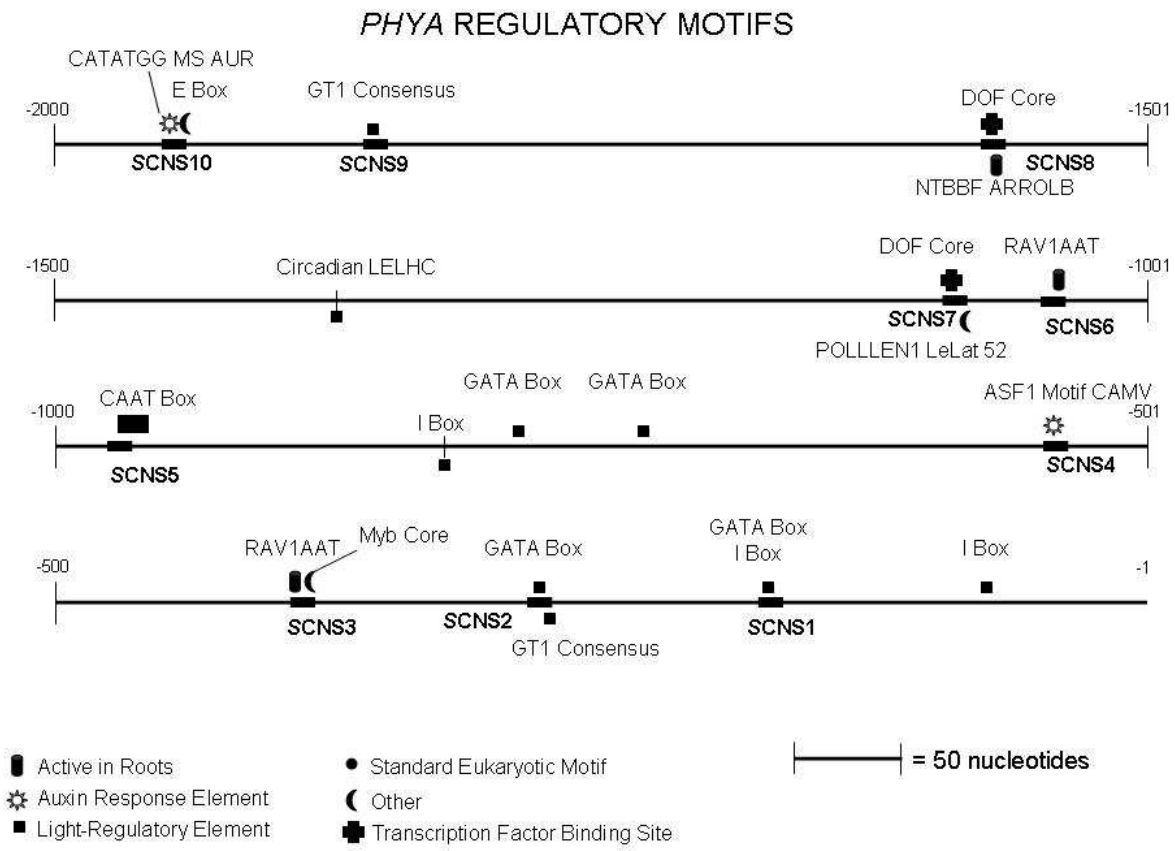


Figure 5.3: Regulatory Motifs Detected Within the Putative Promoter Region of Rice *PHYA*.

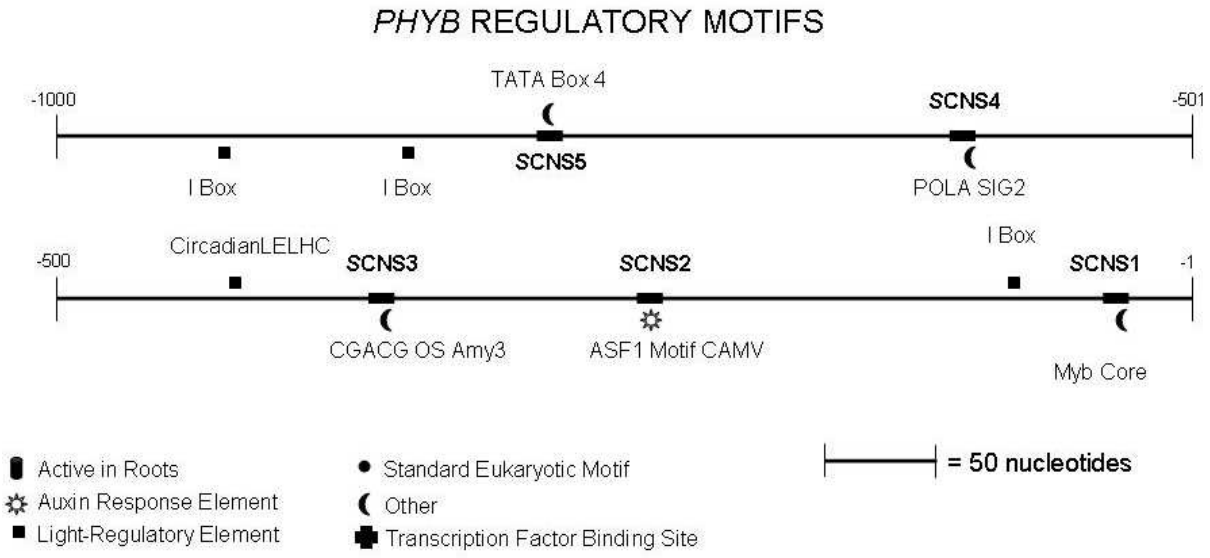


Figure 5.4: Regulatory Motifs Detected Within the Putative Promoter Region of Rice *PHYB*.

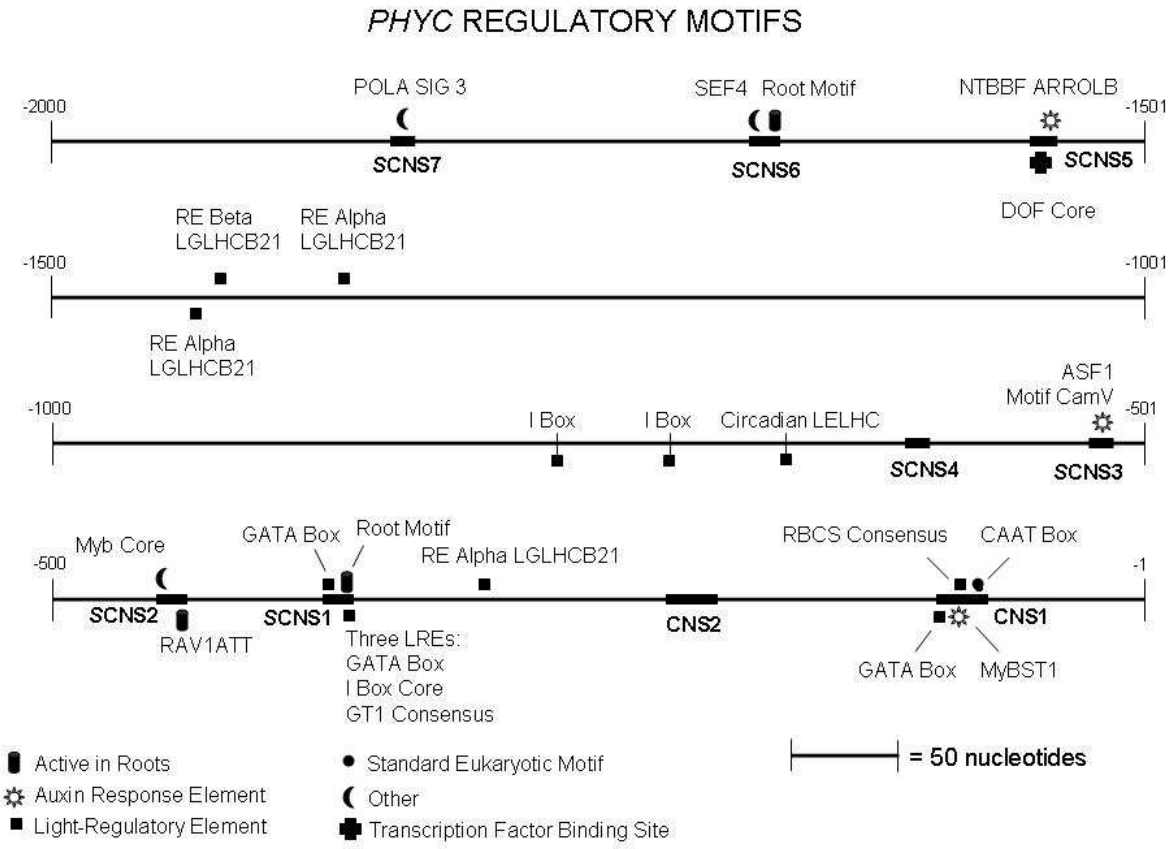


Figure 5.5: Regulatory Motifs Detected Within the Putative Promoter Region of Rice *PHYC*.

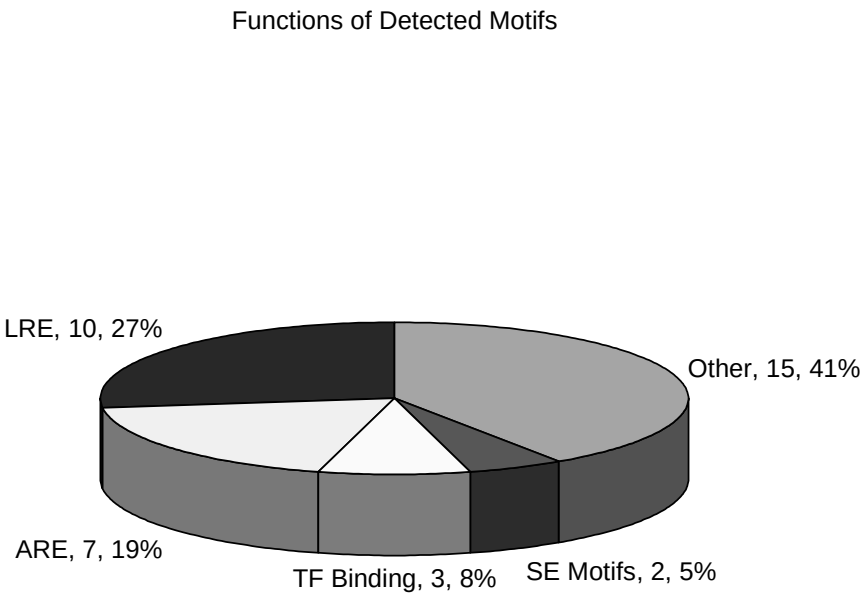


Figure 5.6: Functional Characterization of Detected Regulatory Motifs. The above motifs were detected within the CNSs and *S*CNSs of the rice *PHY* putative promoters. LRE = Light Regulatory Element; ARE = Auxin Regulatory Element; SE = Standard Eukaryotic Regulatory Motif, TF = Transcription Factor binding sites.

References

Abel, S. and Theologis, A. (1996) Early and auxin action. *Plant Physiol*., **111**: 111-117.

Alba, R., Kelmenson, P.M., Cordonnier-Pratt, M.-M., Pratt, L.H. (2000) The phytochrome gene family in tomato and the rapid differential evolution of this family in angiosperms. *Mol. Biol. Evol*., **3**: 362-373.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**:403-410.

Baranowskij, N., Frohberg, C., Prat, S., Willmitzer L. (1994) A novel DNA binding protein with homology to Myb oncoproteins containing only one repeat can function as a transcriptional activator. *EMBO J.*, **13**: 5383-5392.

Basu,D., Dehesh,K., Schneider-Poetsch, H.J., Harrington, S.E., McCouch, S.R. and Quail,P.H. (2000) Rice PHYC gene: structure, expression, map position and evolution. *Plant Mol. Biol.* **44**: 27-42.

Bate, N., Twell, D. (1998) Functional architecture of a late pollen promoter: pollen-specific transcription is developmentally regulated by multiple stage-specific and co-dependent activator elements. *Plant Mol Biol*., **37**: 859-869.

Baumann, K., De Paolis, A., Costantino, P., Gualberti, G. (1999) The DNA binding site of the Dof protein NtBBF1 is essential for tissue-specific and auxin-regulated expression of the rolB oncogene in plants. *Plant Cell*., **11**: 323-334.

Blanchette, M., and Sinha, S. (2001) Separating real motifs from their artifacts. *Bioinformatics* **17**: 30-38.

Bray, N., Dubchak, L., and Pachter, L. (2003) AVID: A Global Alignment Program. *Genome Res*. **13**: 97-102.

Brock, T.G., Ghosheh, N.S., Kaufman, P.B. (1994) Differential sensitivity to indole-3-acetic acid and gibberellic acid following gravistimulation of the leaf sheath pulvini of oat and barley. *Plant Physiol Biochem.*, **32**: 487–491.

Bruce, W.B., Deng, X.W., Quail, P.H. (1991) A negatively acting DNA sequence element mediates phytochrome-directed repression of phyA gene transcription. *EMBO J.*, **10**: 3015-3024.

Bruce, W.B., Quail, P.H. (1990) cis-acting elements involved in photoregulation of an oat phytochrome promoter in rice. *Plant Cell*, **2**: 1081-1089.

Buchel, A.S., Brederode, F.T., Bol, J.F., Linthorst, H.J. (1999) Mutation of GT-1 binding sites in the Pr-1A promoter influences the level of inducible gene expression in vivo. *Plant Mol Biol*., **40**: 387-396.

Bucher, P. (1999) Regulatory elements and expression profiles. *Curr. Opin. Struct. Biol.* **9**: 400-407.

Chattopadhyay, S., Puente, P., Deng, X.W., Wei, N., (1998) Combinatorial interaction of light-responsive elements plays a critical role in determining the response characteristics of light-regulated promoters in Arabidopsis. *Plant J*. **15**: 69-77.

Chen, J., Tang, W.H., Hong, M.M., Wang, Z.Y. (2003) OsBP-73, a rice gene, encodes a novel DNA-binding protein with a SAP-like domain and its genetic interference by double-stranded RNA inhibits rice growth. *Plant Mol Biol*., **52**: 579-590.

Choffnes Inada, D. C., Bashir, A., Lee, C., Thomas, B. C., Ko, C., Goff, S. A., Freeling, M. (2003) Conserved non-coding sequences in the grasses. *Genome Res*. **13**: 2030-2041.

Clua, A., Bottini, R., Brocchi, G.N., Bogino, J., Luna, V., Montaldi, E.R. (1996) Growth habit of *Lotus tenuis* shoots and the influence of photosynthetic photon flux density, sucrose and endogenous levels of gibberellins A-1 and A-3. *Physiol Plant*, 98: 381–388.

Colon-Carmona, A., Chen, D.L., Yeh, K.C., Abel, S (2000) AUX/IAA proteins are phosphorylated by phytochrome in vitro. *Plant Physiol.*, **124**: 1728–1738.

Cox, M.C.H., Millenaar, F.F., de Jong van Berkel, Y.E.M., Peeters, A.J.M., Voesenek, L.A.C.J. (2003) Plant movement: submergence-induced petiole elongation in *Rumex palustris* depends on hyponastic growth. *Plant Physiol.*, **132**: 282–291.

de Freitas, F.A., Yunes, J.A., da Silva, M.J., Arruda, P., Leite, A. (1994) Structural characterization and promoter activity analysis of the gamma-kafirin gene from sorghum. *Mol Gen Genet*., **245**: 177-186.

Degenhardt, J., Tobin, E.M. (1996) A DNA binding activity for one of two closely defined phytochrome regulatory elements in an Lhcb promoter is more abundant in etiolated than in green plants. *Plant Cell*, **8**: 31-41.

Devlin, P.F., Halliday, K.J., Harberd, N.P., Whitelam, G.C. (1996) The rosette habit of Arabidopsis thaliana is dependent upon phytochrome action: novel phytochromes control internode elongation and flowering time. *Plant J.*, **10**: 1127–1134.

Donald, R.G., Cashmore, A.R. (1990) Mutation of either G box or I box sequences profoundly affects expression from the Arabidopsis rbcS-1A promoter.
EMBO J., **9**: 1717-1726.

Dubchak, I., Brudno, M., Loots, G. G., Mayor, C., Pachter, L., Rubin, E.M., and Frazer, K. A. (2000) Active conservation of non-coding sequences revealed by 3-way species comparisons. *Genome Res*. **10**: 1304-1306.

Duret, L. and Bucher, P. (1997) Searching for regulatory elements in human non-coding sequences. *Curr. Opin. Struct. Biol*. **7**: 399-406.

Elmayan, T., Tepfer, M. (1995) Evaluation in tobacco of the organ specificity and strength of the rolD promoter, domain A of the 35S promoter and the 35S2 promoter. *Transgenic Res*., **4**: 388-396.

Feldbrugge, M., Sprenger, M., Hahlbrock, K. and Weisshaar, B. (1997) PcMYB1, a novel plant protein containing a DNA-binding domain with one MYB repeat, interacts in vivo with a light-regulatory promoter unit. *Plant J*., **11**: 1079-1093.

Ficket, J. W., and Hatzigeorgious, A. G. (1997) Eukaryotic promoter recognition. *Genome Res*. **7:** 861-878.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.-I., Postlehwait, J. (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**: 1531-1545.

Gilmartin, P.M., Sarokin, L., Memelink, J. and Chua, N.-H. (1990) Molecular light switches for plant genes. *Plant Cell*, **2**: 369-378.

Green, P. J., Yong, M. H., Cuozzo, M., Kano-Murakami, Y., Silverstein, P., Chua, N.-H. (1988) Binding site requirements for pea nuclear protein factor GT-1 correlate with sequences required for light-dependent transcriptional activation of the rbcS-3A gene. *EMBO J.,* **7**:4035–4044.

Gumucia, D. L., Shelton, D. A., Zhu, W. Millinoff, D., Gray, T., Bock, J.H, Slightom, J.L., Goodman, M. (1996) Evolutionary strategies for the elucidation of *cis* and *trans* factors that regulate the developmental switching programs for the β-like globin genes. *Mol. Phylogenet. Evol*. **5**: 18-32.

Guo, H. and Moose., S. P. (2003) Conserved non-coding sequences among cultivated cereal genomes identity candid regulatory sequence elements and patterns of promoter evolution. *Plant Cell* **15**: 1143-1158.

Ha, S.-B. and An, G. (1988) Identification of upstream regulatory elements involved in the developmental expression of the Arabidopsis thaliana cab1 gene. *Proc. Natl Acad. Sci. USA*, **85**: 8017-8021.

Hangarter, R. P. (1997) Gravity, light, and plant form. *Plant Cell Environ*., **20**: 796-800.

Herdman, M., Coursin, T., Rippka, R., Houmard, J., Tandeau de Marsac, N. (2000) A new appraisal of the prokaryotic origin of eukaryotic phytochromes. *J Mol Evol*., **51**: 205-213.

Higo, K., Ugawa, Y., Iwamoto, M., Korenaga, T. (1999) Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res*. **27**: 297-300.

Holmes, M.G. (1983) Perception of shade. *Philos Trans R Soc Lond B.*, **303**: 503–521.

Hwang, Y.S., Karrer, E.E., Thomas, B.R., Chen, L., Rodriguez, R.L. (1998) Three cis-elements required for rice alpha-amylase Amy3D expression during sugar starvation. *Plant Mol Biol.,* **36**: 331-41.

Joshi, C.P. (1987) Putative polyadenylation signals in nuclear genes of higher plants: a compilation and analysis. *Nucleic Acids Res*., **15**: 9627-9640.

Jürgens, G. (2003) Growing up green: cellular basis of plant development. *Mechanisms of Dev*., **120**: 1395-1406.

Kagaya, Y., Ohmiya, K., Hattori, T. (1999) RAV1, a novel DNA-binding protein, binds to bipartite recognition sequence through two distinct DNA-binding domains uniquely found in higher plants. *Nucleic Acids Res*., **27**: 470-478.

Kaplinsky, N., Barun, D., Penterman, J., Goff, S., Freeling, M. (2002) Utility and distribution of non-coding sequences in the grasses. *Proc. Natl. Acad. Sci., USA* **99**: 6147-6151.

Kasper, S., Matusik, R.J. (2000) Rat probasin: structure and function of an outlier lipocalin. *Biochim Biophys Acta*., **1482**: 249-258.

Katagiri, F., Lam, E., Chua, N.H. (1989) Two tobacco DNA-binding proteins with homology to the nuclear factor CREB. *Nature*, **340**: 727-730.

Klinedinst, S., Pascuzzi, P., Redman, J., Desai, M., Arias, A. (2000) Xenobiotic-stress-activated transcription factor and its cognate target genes are preferentially expressed in root tip meristems. *J. Plant Mol Biol.*, **42**: 679-688.

Kellogg, E. A. (2001) Evolutionary history of the grasses. *Plant Physiol*. **125**: 1198-1205.

Koop, B. F. (1995). Human and rodent DNA sequence comparisons: a mosaic model of genome evolution. *Trends Genet*. **11**: 367-371.

Koornneef, M., Rolff, E., and Spruit, C.J.P. (1980) Genetic control of light-inhibited hypocotyl elongation in *Arabidopsis thaliana. (L.) Heynh. Z Pflanzenphysiol*, **100**: 147–160.

Lam, E., Benfey, P.N., Gilmartin, P.M., Fang, R.X., Chua, N.H. (1989) Site-specific mutations alter in vitro factor binding and change promoter expression pattern in transgenic plants. *Proc Natl Acad Sci, U. S. A.*, **86**: 7890-7894.

Lam, E. Chua, N.-H. (1990) GT-1 binding site confers light responsive expression in transgenic tobacco. *Science*, **248:** 471–474.

Le Gourrierec, J., Li, Y.F., Zhou, D.X. (1999) Transcriptional activation by Arabidopsis GT-1 may be through interaction with TFIIA-TBP-TATA complex. *Plant J*., **18**: 663-668.

Lescot, M., Déhais, P., Thijs, G., Marchal, K., Moreau, Y., Van de Peer, Y., Rouzé, P., Rombauts, S. (2002) PlantCARE, a database of plant *cis*-acting regulatory elements and a portal to tools for in silica analysis of promoter sequences. *Nucleic Acids Res.* **30**: 325-327.

Lessard, P.A., Allen, R.D., Bernier, F., Crispino, J.D., Fujiwara, T., Beachy, R.N. (1991) Multiple nuclear factors interact with upstream sequences of differentially regulated beta-conglycinin genes. *Plant Mol Biol*., **16**: 397-413.

Levy, S., Hannennalli, S., and Workman, C. (2001) Enrichment of regulatory signals in conserved non-coding genomic sequence. *Bioinformatics* **17**: 871-877.

Loots, G. G., Locksley, R. M., Blankespoor, C.M., Wang, Z. E., Miller, W., Rubin, E.M., and Frazer, K.A. (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288**: 136-140.

Maliakal, S.K., McDonnell, K., Dudley, S.A., Schmitt, J. (1999) Effects of red to far-red ratio and plant density on biomass allocation and gas exchange in *Impatiens capensis*. *Int J Plant Sci.,* **160**: 723–733.

Martinez-Garcia, J.F., Huq, E., Quail. P.H. Direct targeting of light signals to a promoter element-bound transcription factor. *Science,* **288**: 859-863.

Miginiac-Maslow, M., Issakidis, E., Lemaire, M., Ruelland, E., Jacquot, J.-P., Decottignies, P. (1997) Light-dependent activation of a NADP-Malate Dehydrogenase: a complex process. *Aust. J. Plant Physiol*., **24**: 529-542.

Millar, A.J. and Kay, S.A. (1996) Integration of circadian and phototransduction pathways in the network controlling CAB gene transcription in Arabidopsis. *Proc. Natl Acad. Sci. USA*, **93**: 15491-15496.

Mindell, D. P., Meyer, A. (2001) Homology evolving. *Trends Ecol. Evol*. **16**: 434-440.

Morgenstern, B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, **15**: 211-218.

Morishige, D. T., Childs, K.L., Moore, L.D., Mullett, J. E. (2002) Targeted analysis of orthologous phytochrome A regions of the sorghum, maize, and rice genomes using comparative gene-island sequencing. *Plant Physiol*. **140**: 1614-1625.

O'Neill, S.D., Kumagai, M.H., Majumdar, A., Huang, N., Sutliff, T.D., Rodriguez, R.L. (1990) The alpha-amylase genes in *Oryza sativa*: characterization of cDNA clones and mRNA expression during seed germination. *Mol Gen Genet.,* **221**: 235-244.

Piechulla, B., Merforth, N., Rudolph, B. (1998) Identification of tomato Lhc promoter regions necessary for circadian expression. *Plant Mol Biol.,* **38**: 655-662.

Prasad, K., Kushalappa, K., Vijayraghavan, U. (2003) Mechanism underlying regulated expression of RFL, a conserved transcription factor, in the developing rice inflorescence. *Mech Dev*., **120**: 491-502.

Praz, V., Perier, R., Bonnard, C., Bucher, P. (2002) The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. *Nucleic Acids Res*. **30**: 322-324.

Puente, P., Wei, N. and Deng, X.W. (1996) Combinational interplay of promoter elements constitutes the minimal determinants for light and developmental control of gene expression in Arabidopsis. *EMBO J,*. **15**: 3732-3743.

Quail, P.H. (1991) Phytochrome: a light-activated molecular switch that regulates plant gene expression. *Ann. Rev. Genet*., **25**: 389-409.

Reed, J.W., Nagpal, P., Poole, D.S., Furuya, M., Chory, J. (1993) Mutations in the gene for the red/far-red light receptor phytochrome B alter cell elongation and physiological responses throughout Arabidopsis development. *Plant Cell,* **5**: 147–157.

Rombauts, S., Florquin, K., Lescot, M., Marchal, K., Rouzé, P., Van de Peer, Y. (2003) Computational approaches to identify promoters and cis-regulatory elements in plant genomes. *Plant Physiol*. **132**: 1162-1176.

Rose, A., Meier, I., Wienand, U. (1999) The tomato I-box binding factor LeMYBI is a member of a novel class of myb-like proteins. *Plant J.,* **20**: 641-652.

Schwartz, S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., Miller, W. (2000) PipMaker--a web server for aligning two genomic DNA sequences. *Genome Res*., **10**: 577-586.

Silverthorne, J. and Tobin, E.M. (1987) Phytochrome regulation of nuclear gene expression. *Bioessays*, **7**: 18-23.

Smith, H., Whitelam, G.C. (1997) The shade avoidance syndrome: multiple responses mediated by multiple phytochromes. *Plant Cell Environ*., **20**: 840–844.

Smith, H. (1992) The ecological functions of the phytochrome family: clues to a transgenic program of crop improvement. *Photochem Photobiol*., **56**: 815–822.

Somers, D. E. Devlin, P. F., Kay, S. A. (1998) Phytochromes and cryptochromes in the entrainment of the Arabidopsis circadian clock. *Science,* 282: 1488-1490.

Sonnhammer, E.L., Durbin, R. (1994) A workbench for large-scale sequence homology analysis. *Comput Appl Biosci*., **10**: 301-307.

Staden, R. (1986) The Current Status and Portability of our Sequence Handling Software. *Nucl. Acids Res*. **14**: 217-231.

Stalberg, K., Ellerstom, M., Ezcurra, I., Ablov, S., Rask, L. (1996) Disruption of an overlapping E-box/ABRE motif abolished high transcription of the napA storage-protein promoter in transgenic *Brassica napus* seeds. *Planta*., **199**: 515-519.

Steindler, C., Matteucci, A., Sessa, G., Weimar, T., Ohgishi, M., Aoyama, T., Morelli, G., Ruberti, I. (1999) Shade avoidance responses are mediated by the ATHB-2 HD-Zip protein, a negative regulator of gene expression. *Development,* **126**: 4235–4245.

Somers, D.E., Sharrock, R.A., Tepperman, J.M., Quail, P.H. (1991) The Hy3 long hypocotyl mutant of Arabidopsis is deficient in phytochrome-B. *Plant Cell,* **3**: 1263–1274.

Sun, L. and Tobin, E.M. (1990) Phytochrome-regulated expression of genes encoding light-harvesting chlorophyll a/b-protein in two long hypocotyl mutants and wild type plants of Arabidopsis thaliana. *Photochem. Photobiol*., **15**: 51-56.

Swarup, R., Parry, G., Graham, N., Allen, T., Bennett, M. (2002) Auxin cross-talk: integration of signaling pathways to control plant development. *Plant Mol Biol.,* **49**: 411–426.

Tagle, D. A., Koop, B. F., Goodman, M., Slightom, J.L., Hess, D.L., Jones, R.T. (1988) Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*): nucleotide and amino acid sequences, developmental regulation, and phylogenetic footprints. *J. Mol. Biol*. **203**: 439-455.

Tahir, M., Kanegae,H., and Takano,M. (2003) Rice phytochrome genes. Published Only in GenBank Database.

Tatusova, T. A. and Madden, T. L. (1999) BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. *FEMS Microbiol. Lett*. **174**: 247-250.

Teakle, G.R., Manfield, I.W., Graham, J.F., Gilmartin, P.M. (2002) *Arabidopsis thaliana* GATA factors: organisation, expression and DNA-binding characteristics. *Plant Mol Biol.,* **50**: 43-57.

Terzaghi, W.B. and Cashmore, A.R. (1995) Light-regulated transcription. Ann. Rev. Plant Physiol. *Plant Mol. Biol*., **46**: 445-474.

Tobin, E. M., Kehoe, D. M. (1994) Phytochrome regulated gene expression. *Semin Cell Biol.,* **5**: 335–346.

Urao, T., Yamaguchi-Shinozaki, K., Urao, S., Shinozaki, K. (1993) An Arabidopsis myb homolog is induced by dehydration stress and its gene product binds to the conserved MYB recognition sequence. *Plant Cell.,* **5**: 1529-1539.

Vandenbussche, F., Vriezen, W.H., Smalle, J., Laarhoven, L.J., Harren, F.J., Van Der Straeten, D. Ethylene and auxin control the Arabidopsis response to decreased light intensity. *Plant Physiol*.,**133**: 517-527.

Villain, P., Mache, R., Zhou, D.X. (1996) The mechanism of GT element-mediated cell type-specific transcriptional control. *J Biol Chem*., **271**: 32593-32598.

Wasserman, W. W., Palumbo, M., Thompson, W., Fickett, J.W., Lawrence, C. E. (2000) Human-mouse genome comparisons to locate regulatory sites. *Nat. Genet*. **26**: 225-228.

White,G.M., Hamblin,M.T. and Kresovich,S. (2004) Molecular evolution of the phytochrome gene family in sorghum: changing rates of synonymous and replacement evolution. *Mol. Biol. Evol.* **21**: 716-723.

Wingender, E., Dietze, P., Karas, H., Knuppel, R. (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res. **24**: 238-241.

Xu, N., Hagen, G., Guilfoyle, T. (1997) Multiple auxin response modules in the soybean SAUR 15A promoter. *Plant Sci.,* **126**: 193-201.

Yanagisawa, S., Schmidt, R.J. (1999) Diversity and similarity among recognition sequences of Dof transcription.factors. *Plant J.*, **17**: 209-214.

Zhu, J., Liu, J.S., Lawrence, C.E. (1998) Bayesian adaptive sequence alignment algorithms. *Bioinformatics*, **14**: 25-39.

CHAPTER 6

GLOBAL CONCLUSION

The overall purpose of this study was to investigate the phytochrome gene family in rice, a model monocot. The project was organized into three main goals:

(I) Develop a system that can handle the appropriate level of detail necessary for an in-depth and accurate annotation of *PHY*[1]-containing sequences.

(II) Sequence the *PHY* and their flanking DNA in rice and extensively annotate these sequences to determine what genomic features (genes, non-coding sequences, repeat sequences, etc.), if any, are conserved between these three regions of the rice genome and *PHY*-containing regions of any publicly available plant genome.

(III) Analyze the putative upstream promoter regions of the *PHY* and the promoters of sorghum *PHY* to identify conserved non-coding sequences between the putative promoters. Identify regulatory motif sequences within conserved regions detected between the sorghum/rice *PHY* orthologs. Also, identify any light-regulatory motifs present in the sequence flanking the conserved regions in the rice *PHY* promoters.

Development of a System for Detailed Annotation

The annotation method described here was designed specifically for a small- to medium-throughput sequencing project that requires detailed and customized annotation. Moreover, this method was developed in an attempt to provide a more appropriate system for a user seeking to investigate a specific topic that pertains to a specific portion of a genome.

A relational database was created to store gene prediction data from five well-known gene prediction programs (FgeneSH, GeneMark, Genscan, and two versions of RiceHMM). Scripts were written to reformat the output files of each prediction program, upload the results

[1] *PHY* = phytochrome gene, PHY = phytochrome protein. A specific *PHY* or PHY is designated by the letter(s) of that *PHY*/PHY or *PHY*/PHY subfamily.

into the database schema, and consequently reformat the results from the database into a format that could be read by Artemis, an annotation and genomic sequence viewing tool. Artemis was then used to create fastA files of the predicted gene sequences, sequences of open reading frames greater than some length defined by the user, and any other sequence that warranted further investigation. These sequences were then compared to databases of known sequences and the sequence homology results were parsed by perl scripts and uploaded to a second database schema. The results of the sequence homology searches were then reformatted for Artemis by perl scripts.

The use of the database schemas imparts flexibility and data storage, retrieval, and query capabilities not offered by other currently available annotation systems of this general type. The combination of multiple *ab initio* prediction programs and homology searches increased the accuracy of gene finding within the sequences. Also, displaying the prediction and sequence comparison data in an interactive graphics-based viewer permitted the curation and organization of all available data to facilitate and fine-tune the annotations. The flexibility of this method is evident in the potential addition of more customized gene prediction programs or sequence homology searches by the annotator. The use of two database schemas further facilitates storage of the prediction data in a format that can be easily manipulated in order to customize it for other analyses.

Sequencing and Annotating PHY*-containing BACs and Comparisons of* PHY*-containing Genomic Sequences*

The *PHY* family and genomic sequence surrounding each member was successfully sequenced by a shotgun approach. These BAC sequences were confirmed by restriction digest analysis and deposited into GenBank for availability to the genome community at large.

The three BAC sequences were extensively annotated by manually curating gene prediction and sequence similarity data generated by running multiple prediction algorithms and performing many comparisons to databases of known genes, proteins, and protein domains. It was determined that employing multiple prediction programs greatly increases the overall accuracy of gene prediction and that of the prediction algorithms used here, FgeneSH was the most accurate in predicting genes in this region of the rice genome. The genes surrounding *PHY* were found not to display any discernable functional relationship to one anther.

These *PHY*-containing genomic sequences of rice were compared to each other and to other plant *PHY*-containing sequences to investigate any possible mechanisms of *PHY* family expansion in plants. Although 21 comparisons were performed with five different methods, no evidence or information regarding the mechanisms of *PHY*-family expansion could be uncovered. The lack of detectable sequence similarity within these areas of the rice, Arabidopsis, and sorghum genomes is not surprising given the ancient nature of the phytochrome gene duplications.

Analysis of Conserved Sequences in the Putative Promoters of PHY

Employing a comparative genomics approach to reveal patterns of noncoding sequence conservation and evolution, particularly in promoter regions, can offer valuable insights into the

complex regulation of gene expression in plants. Putative upstream promoter regions of rice *PHY* and the promoters of sorghum *PHY* were analyzed to identify conserved non-coding sequences between their putative promoters. Also, regulatory motifs within the conserved regions detected between the sorghum/rice *PHY* orthologs and light-regulatory motif elements flanking the conserved regions in the rice *PHY* promoters were identified. Although the lack of both putative and experimentally defined promoter sequences from *PHY* and other light-inducible genes from plants made it difficult to generate an abundance of sequence for meaningful comparisons. Also, the short nature of the conserved sequences present between the orthologous and paralogous *PHY* also made it impossible to statistically verify their significance.

It was found that BL2SEQ, a local alignment algorithm, detects more CNSs and *S*CNSs than *AVID*, a global alignment tool. This is most likely due to the restriction inherent to *AVID* that sequences of similarity must be present in the same order and conservation in to be considered biologically relevant. Regardless of the method of comparison employed, however, the function of the proteins encoded by the members of a gene pair effects the number and amount of conservation detected more so than the genetic relationship of the members. Regulatory motifs of multiple functions can be found within the conserved regions of the *PHY* putative promoters and LREs are the most abundantly detected type of motif within the regions of conservation. Auxin response elements are the second most frequently identified regulatory motif in *PHY* promoters, although it is unknown if this is due to a relationship between auxin-response and photo-regulation or because the auxin response elements are so well characterized.

APPENDICES

Appendix A: Creation Script for Gene Prediction and Analysis Schema. This schema and script was designed and written by S. Khosla, A. Eastman, and M. Shah.

```
create table PROGRAM
(
        PRGM_ID                 NUMBER(5)       PRIMARY KEY,
        PRGM_NAME               VARCHAR2(15) NOT NULL, --Exs: Genscan, blastx
        VERSION                 VARCHAR2(5),
        MATRIX_FILE             VARCHAR2(60),
        DATE_ACQUIRED           VARCHAR2(6)  NOT NULL,
        COMMENT                 VARCHAR2(500)
);

create table TEMPLATE
(
        TEMPLATE_ID             NUMBER(5)       PRIMARY KEY,
        TEMPLATE_NAME           VARCHAR2(20) NOT NULL,
        TEMPLATE_TYPE           VARCHAR2(20),           --BAC, etc
        TEMPLATE_GT_COMBINE_CODE
        TEMPLATE_ALIAS          VARCHAR2(30),
        TEMPLATE_REFERENCE      VARCHAR2(30),           --template common name
        ASSEMBLY_DATE           VARCHAR2(6),
        ACE_FILE_NAME           VARCHAR2(30),
        COMMENT                 VARCHAR2(500)
);

create table GENSCAN
(
        PROGRAM_ID              NUMBER(5)       NOT NULL,
        GENE_NUM                NUMBER(5)       NOT NULL,
        TEMPLATE_ID             NUMBER(5)       NOT NULL,
        FEATURE_START           NUMBER(12)      NOT NULL,
        FRAME                   NUMBER(1),
        NET_PHASE_EXON          NUMBER(1),
        INIT_SIG_SPLICE_SCORE,
        COD_REG_SCORE,
        PROBABILITY             NUMBER(5),      --decimal format
        EXON_SCORE,
        PRIMARY KEY (PROGRAM_ID, TEMPLATE_ID, FEATURE_START),
        FOREIGN KEY (PROGRAM_ID, TEMPLATE_ID)
);

create table GENEMARK
(
        PROGRAM_ID              NUMBER(5)       NOT NULL,
        TEMPLATE_ID             NUMBER(5)       NOT NULL,
        FEATURE_START           NUMBER(12)      NOT NULL,
        UPDATE_DATE             VARCHAR2(6)  NOT NULL,
        START_FRAME             NUMBER(1),
        END_FRAME               NUMBER(1),
        PRIMARY KEY (PROGRAM_ID, TEMPLATE_ID, FEATURE_START, UPDATE_DATE),
        FOREIGN KEY (PROGRAM_ID, TEMPLATE_ID)
);
```

```
create table RICEHMM
(
        PROGRAM_ID                  NUMBER(5)     NOT NULL,
        TEMPLATE_ID                 NUMBER(5)     NOT NULL,
        FEATURE_START               NUMBER(12)    NOT NULL,
        UPDATE_DATE                 VARCHAR2(6)   NOT NULL,
        INDEX                       ,
        FRAME                       NUMBER(1),
        PRIMARY KEY (PROGRAM_ID, TEMPLATE_ID, FEATURE_START, UPDATE_DATE),
        FOREIGN KEY (PROGRAM_ID, TEMPLATE_ID)
);

create table FGENESH
(
        PROGRAM_ID                  NUMBER(5)     NOT NULL,
        TEMPLATE_ID                 NUMBER(5)     NOT NULL,
        FEATURE_START               NUMBER(12)    NOT NULL,
        UPDATE_DATE                 VARCHAR2(6)   NOT NULL,
        SCORE                       ,
        PRIMARY KEY (PROGRAM_ID, TEMPLATE_ID, FEATURE_START, UPDATE_DATE),
        FOREIGN KEY (PROGRAM_ID, TEMPLATE_ID)
);

create table ANNOTATION
(
        ANNOTATION_ID               NUMBER(5)     PRIMARY KEY,
        PROGRAM_ID                  NUMBER(5)     NOT NULL,
        TEMPLATE_ID                 NUMBER(5)     NOT NULL,
        FEATURE_START               NUMBER(12)    NOT NULL,
        UPDATE_DATE                 VARCHAR2(6)   NOT NULL,
        GENE_NUM                    NUMBER(6),
        FEATURE_NUM                 NUMBER(6),
        FEATURE_TYPE                VARCHAR(10),
        FEATURE_STOP                NUMBER(12),
        PROG_COMBO_ID               NUMBER(3), --way to know what combination of programs
                                      result in this feature being annotated,
                                      manual=0
        MAN_FEATURE_NUM             NUMBER(3), --allows you to assign a number to a manual
                                      annotation
        FINAL_TAG                   NUMBER(1), --Is this your final annotation assignment?
                                      1=yes and 0=no?
        FINAL_DETECTION_ID NUMBER(3),
        UNIQUE (PROGRAM_ID, TEMPLATE_ID, FEATURE_START, UPDATE_DATE),
        FOREIGN KEY (PROGRAM_ID, TEMPLATE_ID)
);
```

Appendix B: Perl Script artparse.pl. This script was designed and implemented by A. Eastman and D. Kolychev.

```
#! /usr/bin/perl
use URI::URL;
use Tk;
use Tk::DialogBox;
use Tk::NoteBook;
use Tk::LabEntry;
use LWP::Simple;
use String::Similarity;
use POSIX;
use threads;
use threads::shared;

use lib ("/home/peela");
use uga_perl::util::util_oradbi
        qw(getConnection);

#use strict;     # Always!
#$^W++;          # Turn on warnings

##### Global Configuration Structures
####
$ENV{ORACLE_HOME}= "/oracle/ora901";
my $DEBUG = 1;  # turn off (0) if you don't want to see internals
my $mainwindow;
my $f;          # dialog window
my $f1;
our $grun : shared;

our %trans : shared;
$trans= {    _scanr => '',
             _filename => '',
             _tid => '',
        _tname     => '',
        _ttype    => '',
        _tgtcode  => '',
        _talias   => '',
        _tref   => '',
        _assemblydate   => '',
        _acename    => '',
            _comments => '',
             };
```

```
# Begin main program
MAIN: {
  debug("+MAIN");

  # Attempt to load the master.pcg configuration file

  $mainwindow = MainWindow->new();
  $mainwindow->title("Gene Prediction Parsing");
  #$mainwindow->minsize(qw(300 150));
  #$mainwindow->maxsize(qw(300 150));
  #$mainwindow->geometry('+250+150');
  my $left   = $mainwindow->Frame->grid(-row => 1,
                                -col => 0,
                                -sticky => 'nw');
  my $right  = $mainwindow->Frame->grid(-row => 1,
                                -col => 1,
                                -sticky => 'nw');
  my $bottom = $mainwindow->Frame->grid(-row => 2,
                                -col => 0,
                                -columnspan => 3,
                                -sticky => 'nw');
  ### Left panel, input functionality

    $left->LabEntry(-label => "filename
(templateName_aceFileName_assemblyDate_programID)",
     -labelPack => [-side => "right", -anchor => "w"],
     -width => 20,
     -textvariable => \$trans->{_filename})->pack(-side => "top",
                                       -anchor => "nw");
  $left->LabEntry(-label => "template type",
     -labelPack => [-side => "right", -anchor => "w"],
     -width => 20,
     -textvariable => \$trans->{_ttype})->pack(-side => "top",
                                       -anchor => "nw");
  $left->LabEntry(-label => "template GT_COMBINE_CODE",
     -labelPack => [-side => "right", -anchor => "w"],
     -width => 20,
     -textvariable => \$trans->{_gtcode})->pack(-side => "top",
                                       -anchor => "nw");
  $left->LabEntry(-label => "template alias",
     -labelPack => [-side => "right", -anchor => "w"],
     -width => 20,
     -textvariable => \$trans->{_talias})->pack(-side => "top",
                                       -anchor => "nw");
  $left->LabEntry(-label => "template reference",
     -labelPack => [-side => "right", -anchor => "w"],
```

```
    -width => 20,
    -textvariable => \$trans->{_tref})->pack(-side => "top",
                                    -anchor => "nw");
  $left->LabEntry(-label => "comments",
    -labelPack => [-side => "right", -anchor => "w"],
    -width => 20,
    -textvariable => \$trans->{_comments})->pack(-side => "top",
                                    -anchor => "nw");
  $bottom->Button(-text => 'RUN',
                -command => \&processTransaction )->
                grid(qw/-row 3 -column 0 -sticky nesw/);

  MainLoop();   # Start the event processing

  # Will never get here
  debug ("-MAIN");
}
sub processTransaction {
  #print "running is $grun.\n";
  if(!$grun){
    $grun=1;
    #print "running 1 is $grun.\n";
    $thr= threads->new(\&parse);
  }
  else{
    $f1 = $mainwindow->DialogBox(-title => "Process running: unable to continue",
                          -buttons => ["Cancel"]);
    my $result = $f1->Show;
    debug("already running\n");
}
}
sub parse{
debug("+PARSE");
my $DB_Type="test";
my $Oracle_ID="version2_GP";
my $Oracle_Passwd="version2_GP";
my $dbh = getConnection($DB_Type, $Oracle_ID, $Oracle_Passwd);

my @fields=split("_", $trans->{_filename});
$trans->{_tid}=$fields[0]."_".$fields[1]."_".$fields[2];
$trans->{_tname}=@fields[0];
$trans->{_acename}=@fields[1];
$trans->{_assemblydate}=@fields[2];
$trans->{_scanr}=@fields[3];
my $sth = $dbh->prepare("select program_name from v2_programs where program_id=$trans-
>{_scanr}");
```

```
$sth->execute();
my @ary = $sth->fetchrow_array();
my $pname=shift @ary;
debug("pname is $pname");
$sth->finish();

my $sth = $dbh->prepare("select update_date from v2_programs where program_id=$trans-
>{_scanr}");
$sth->execute();
my @ary = $sth->fetchrow_array();
my $ud=shift @ary;
debug("update_date is $ud");
$sth->finish();

$sth = $dbh->prepare("select template_id from v2_template where template_id=\'$trans-
>{_tid}\'");
$sth->execute();
@ary = $sth->fetchrow_array();
my $exists=shift @ary;
$sth->finish();
if(!$exists){
    $sth = $dbh->prepare("insert into version2_GP.v2_template (template_id, template_name,
template_type, template_gt_combine_code, template_alias, template_reference, assembly_date,
ace_file_name, comments) values (\'$trans->{_tid}\', \'$trans->{_tname}\', \'$trans->{_ttype}\',
\'$trans->{_gtcode}\', \'$trans->{_talias}\', \'$trans->{_tref}\', \'$trans->{_assemblydate}\',
\'$trans->{_acename}\', \'$trans->{_comments}\')");
    $sth->execute();
    $sth->finish();
}
$dbh->disconnect;

my $genescancommand="perl ".$pname."2ora_v1.txt ".$trans->{_tid}." ".$trans-
>{_tname}."_".$trans->{_acename}."_".$trans->{_assemblydate}."_".$trans->{_scanr}."
".$trans->{_scanr}." ".$ud;
system("$genescancommand");

debug("-PARSE");
print("$trans->{_filename} parsing process COMPLETE\n");
$grun=0;
}
sub debug {
  my @msg = shift;
  print @msg, "\n" if $DEBUG;
}
```

Appendix C: Perl Script FGeneSH2Ora.pl. This script was designed and implemented by A. Eastman and D. Kolychev.

```
#!/usr/bin/perl -w
#use strict;
use LWP::Simple;
use lib ("/home/peela");
use uga_perl::util::util_oradbi
       qw(getConnection);
use Bio::SeqIO;
my @filearr;
my %maxexon;
my $currgene=-1;

#------------------------------ Main ------------------------------------------------

## ---- Check Command Line Argument ----

if (@ARGV != 4) {
  die " Invalid Arguments! \n Usage: <scriptname>
<template_id><genemark_output><program_id><update_date>\n";
}
Insert2Ora($ARGV[0],$ARGV[1],$ARGV[2],$ARGV[3]);

#------------------------------ End of Main -----------------------------------------

#------------------------------ Insert2Ora ------------------------------------------
sub Insert2Ora {
my $readline="";
my($tid,$GScanOutFile,$pid,$update_date) = @_ ;
my @bacfilearr=split('_',$tid);
my $bacfile=$tid[0]."_".$tid[1];
$inbac=Bio::SeqIO->new('-file' => $bacfile, '-format'=>'Fasta');
my $bac=$inbac->next_seq();
open (GSCANOUT, $GScanOutFile) or die "Cannot open the input file: $GScanOutFile \n";
while ($readline = <GSCANOUT>) {
   chomp($readline);
   push (@filearr, $readline);
   my @GeneExon;
   my $Gene;
   my $Exon;
   my $UniqueID;
   my $Type;
   my $Strand;
   my $ExonB;
   my $ExonE;
```

```
my $ExonL;
my $Fr;
my $Ph;
my $IAc;
my $DoT;
my $CodRg;
my $P;
my $Tscr;
my $ORFB;
my $ORFE;
my $Score;

if ($readline =~ /\sTSS\s|\sCDSf\s|\sCDSl\s|\sCDSi\s|\sCDSo\s|\sPolA\s/) {

        my @fields = split (' ',$readline);

  if ($readline =~ /CDSf|CDSl|CDSi|CDSo/) {
    $Gene   = $fields[0];
    $Strand = $fields[1];
    $Exon   = $fields[2];
    $Type   = $fields[3];
    $ExonB  = $fields[4];
    $ExonE  = $fields[6];
    $Score  = $fields[7];
    $ORFB   = $fields[8];
        $ORFE   = $fields[10];
    $ExonL  = $fields[11];
   }
   else {
        $Gene   = $fields[0];
    $Strand = $fields[1];
    $Exon   = "NULL";
    $Type   = $fields[2];
    $ExonB  = $fields[3];
    $ExonE  = "NULL";
    $Score  = $fields[4];
    $ORFB   = "NULL";
    $ORFE   = "NULL";
    $ExonL  = "NULL";
   }
        if ($Type eq "CDSf"){
       .....$Type = "Initial";
    }
        elsif ($Type eq "CDSl"){
       .....$Type = "Terminal";
       ..... }
```

```
            elsif ($Type eq "CDSi"){
            .....$Type = "Internal";
            ..... }
            elsif ($Type eq "CDSo"){
            ..... $Type = "Single";
            }

      if ($Strand eq "+") {
            $Strand = "plus";
      }
      elsif ($Strand eq "-") {
            $Strand = "minus";
      }

      if($Strand eq "plus"){
            my $DB_Type="test";
            my $Oracle_ID="version2_GP";
            my $Oracle_Passwd="version2_GP";
            my $dbh = getConnection($DB_Type, $Oracle_ID, $Oracle_Passwd);

            my $sth = $dbh->prepare("insert into version2_GP.v2_fgenesh (gene_num,
template_id, feature_start, score, program_id, update_date) values ($Gene, \'$tid\', $ExonB,
$Score, $pid, \'$update_date\')");
            $sth->execute();
            $sth->finish();

            $sth = $dbh->prepare('select annotation_id.nextval from dual');
            $sth->execute();
            my @ary = $sth->fetchrow_array();
            my $aid=shift @ary;
            $sth->finish();

            my $ntexonseq=getntexonseq($bac, $ExonB, $ExonL);
            my $aaexonseq=getaaexonseq($bac, $ExonB, $ExonL);

            $sth = $dbh->prepare("insert into version2_GP.v2_annotation (annotation_id,
pgrm_gene_num, template_id, strand, feature_type, feature_start, feature_stop, feature_length,
exon_num, program_id, update_date ) values ($aid, $Gene, \'$tid\', \'$Strand\', \'$Type\',
$ExonB,  $ExonE, $ExonL, $Exon, $pid, \'$update_date\')");
            $sth->execute();
            $sth->finish();
            if(genecomplete($Gene){
            ..... print("update gene $Gene in all exons\n");
            }
            $dbh->disconnect;
          }
```

```
            elsif($Strand eq "minus" && $readline =~ /CDSf|CDSl|CDSi|CDSo/){
              if(!$maxexon->{$Gene}){
              ..... $maxexon->{$Gene}=1;
              }
              else{
              ..... $maxexon->{$Gene}++;
              }
            }
}
}
}
close GSCANOUT;
foreach $readline(@filearr){
   my @GeneExon;
   my $Gene;
   my $Exon;
   my $UniqueID;
   my $Type;
   my $Strand;
   my $ExonB;
   my $ExonE;
   my $ExonL;
   my $Fr;
   my $Ph;
   my $IAc;
   my $DoT;
   my $CodRg;
   my $P;
   my $Tscr;
   my $ORFB;
   my $ORFE;
   my $Score;
   if ($readline =~ /\sTSS\s|\sCDSf\s|\sCDSl\s|\sCDSi\s|\sCDSo\s|\sPolA\s/) {

            my @fields = split (' ',$readline);
     if ($readline =~ /CDSf|CDSl|CDSi|CDSo/) {
       $Gene   = $fields[0];
       $Strand = $fields[1];
       $Exon   = $fields[2];
       $Type   = $fields[3];
       $ExonB  = $fields[4];
       $ExonE  = $fields[6];
       $Score  = $fields[7];
       $ORFB   = $fields[8];
       $ORFE   = $fields[10];
       $ExonL  = $fields[11];
```

```
        }
        else {
             $Gene   = $fields[0];
         $Strand = $fields[1];
         $Exon   = "NULL";
         $Type   = $fields[2];
         $ExonB  = $fields[3];
         $ExonE  = "NULL";
         $Score  = $fields[4];
         $ORFB   = "NULL";
         $ORFE   = "NULL";
         $ExonL  = "NULL";
        }
             if ($Type eq "CDSf"){
            .....$Type = "Initial";
         }
             elsif ($Type eq "CDSl"){
            .....$Type = "Terminal";
            ..... }
             elsif ($Type eq "CDSi"){
            .....$Type = "Internal";
            ..... }
             elsif ($Type eq "CDSo"){
            ..... $Type = "Single";
             }

      if ($Strand eq "+") {
             $Strand = "plus";
      }
      elsif ($Strand eq "-") {
             $Strand = "minus";
      }
      if($Strand eq "minus"){
             my $DB_Type="test";
             my $Oracle_ID="version2_GP";
             my $Oracle_Passwd="version2_GP";
             my $dbh = getConnection($DB_Type, $Oracle_ID, $Oracle_Passwd);
             my $sth = $dbh->prepare("insert into version2_GP.v2_fgenesh (gene_num,
template_id, feature_start, score, program_id, update_date) values ($Gene, \'$tid\', $ExonB,
$Score, $pid, \'$update_date\')");
             $sth->execute();
             $sth->finish();

             $sth = $dbh->prepare('select annotation_id.nextval from dual');
             $sth->execute();
             my @ary = $sth->fetchrow_array();
```

```
            my $aid=shift @ary;
            $sth->finish();

            if($readline =~ /CDSf|CDSl|CDSi|CDSo/){
          .....$Exon=$maxexon->{$Gene};
          .....$maxexon->{$Gene}--;
            }
            $sth = $dbh->prepare("insert into version2_GP.v2_annotation (annotation_id,
pgrm_gene_num, template_id, strand, feature_type, feature_start, feature_stop, feature_length,
exon_num, program_id, update_date ) values ($aid, $Gene, \'$tid\', \'$Strand\', \'$Type\',
$ExonB,  $ExonE, $ExonL, $Exon, $pid, \'$update_date\')");
            $sth->execute();
            $sth->finish();

            $dbh->disconnect;
          }  .....
}
}
sub genecomplete{
   my $gene=shift @_;
   if($currgene==-1){
         $currgene=$gene;
   }
   if($gene!=$currgene){
         $currgene=$gene;
         return 1;
   }
   else{
         return 0;
   }
}
sub getntexonseq{
   my ($bac, $start, $end)=@_;
   my $ntexon   = $bac->subseq($start,$end);
   return ntexon;
}

sub getaaexonseq{
   my ($bac, $start, $len)=@_;
   my $ntexon   = $bac->subseq($start,$end);
   return dna2peptide($ntexon);
}
```

Appendix D: Perl Script GeneMark2Ora.pl. This script was designed and implemented by A. Eastman and D. Kolychev.

```
#!/usr/bin/perl -w
#use strict;
use LWP::Simple;
use lib ("/home/peela");
use uga_perl::util::util_oradbi
       qw(getConnection);

my %maxexonf;

#------------------------------ Main ------------------------------------------------

## ---- Check Command Line Argument ----

if (@ARGV != 4) {
  die " Invalid Arguments! \n Usage: <scriptname>
<template_id><output_file><program_id><update_date>\n";
}

Insert2Ora($ARGV[0],$ARGV[1],$ARGV[2],$ARGV[3]);
#------------------------------ End of Main ----------------------------------------

#------------------------------ Insert2Ora ----------------------------------------

sub Insert2Ora {

my $readline="";
my($tid,$GScanOutFile,$pid,$update_date) = @_ ;
open (GSCANOUT, $GScanOutFile) or die "Cannot open the input file: $GScanOutFile \n";

while ($readline = <GSCANOUT>) {

   chomp($readline);

   my @GeneExon;
   my $Gene;
   my $Exon;
   my $UniqueID;
   my $Type;
   my $Strand;
   my $ExonB;
   my $ExonE;
   my $ExonL;
   my $Fr;
```

```
    my $Ph;
    my $IAc;
    my $DoT;
    my $CodRg;
    my $P;
    my $Tscr;
    my $StartF;
    my $EndF;

#########
    if ($readline =~ /Terminal|Initial|Internal|Single/)
            my @fields = split (' ',$readline);

        $Gene   = $fields[0];
        $Exon   = $fields[1];
        $Strand = $fields[2];
        $Type   = $fields[3];
        $ExonB  = $fields[4];
        $ExonE  = $fields[5];
        $ExonL  = $fields[6];
        $StartF = $fields[7];
        $EndF   = $fields[8];

        $UniqueID = "$Gene.$Exon";

        #print (OUT
"$UniqueID,$Gene,$Exon,$Strand,$Type,$ExonB,$ExonE,$ExonL,$StartF,$EndF,");

#########

      if ($Strand eq "+") {
            $Strand = "plus";
      }
      elsif ($Strand eq "-") {
            $Strand = "minus";
      }

      if($Type eq "Init"){
            .....$Type="Initial";
      }
      elsif($Type eq "Term"){
            .....$Type="Terminal";
      }
      elsif($Type eq "Intr"){
            .....  $Type="Internal";
      }
```

```
      if ($readline =~ /Terminal|Initial|Internal|Single/) {
            if(!$maxexonf->{$Gene}){
           .....$maxexonf->{$Gene}=1;
            }
            else{
           .....$maxexonf->{$Gene}++;
            }
          }

}
}
close GSCANOUT;

open (GSCANOUT, $GScanOutFile) or die "Cannot open the input file: $GScanOutFile \n";

while ($readline = <GSCANOUT>) {

   chomp($readline);

   my @GeneExon;
   my $Gene;
   my $Exon;
   my $UniqueID;
   my $Type;
   my $Strand;
   my $ExonB;
   my $ExonE;
   my $ExonL;
   my $Fr;
   my $Ph;
   my $IAc;
   my $DoT;
   my $CodRg;
   my $P;
   my $Tscr;
   my $StartF;
   my $EndF;

#########
   if ($readline =~ /Terminal|Initial|Internal|Single/) {

         my @fields = split (' ',$readline);

      $Gene   = $fields[0];
      $Exon   = $fields[1];
      $Strand = $fields[2];
```

```
        $Type   = $fields[3];
        $ExonB  = $fields[4];
        $ExonE  = $fields[5];
        $ExonL  = $fields[6];
        $StartF = $fields[7];
        $EndF   = $fields[8];

        $UniqueID = "$Gene.$Exon";

        #print (OUT
"$UniqueID,$Gene,$Exon,$Strand,$Type,$ExonB,$ExonE,$ExonL,$StartF,$EndF,");
#########

      if ($Strand eq "+") {
            $Strand = "plus";
      }
      elsif ($Strand eq "-") {
            $Strand = "minus";
      }

      if($Type eq "Init"){
            .....$Type="Initial";
      }
      elsif($Type eq "Term"){
            .....$Type="Terminal";
      }
      elsif($Type eq "Intr"){
            .....  $Type="Internal";
      }

      if ($readline =~ /Terminal|Initial|Internal|Single/) {
           if($maxexonf->{$Gene}==1){
            .....$Type="Single";
            ..... $Exon=1;
           }
           }
my $DB_Type="test";
my $Oracle_ID="version2_GP";
my $Oracle_Passwd="version2_GP";
my $dbh = getConnection($DB_Type, $Oracle_ID, $Oracle_Passwd);

my $sth = $dbh->prepare("insert into version2_GP.v2_genemark (gene_num, template_id,
feature_start, start_frame, end_frame, program_id, update_date) values ($Gene, \'$tid\', $ExonB,
$StartF, $EndF, $pid, \'$update_date\')");
$sth->execute();
$sth->finish();
```

```
$sth = $dbh->prepare('select annotation_id.nextval from dual');
$sth->execute();
my @ary = $sth->fetchrow_array();
my $aid=shift @ary;
$sth->finish();

$sth = $dbh->prepare("insert into version2_GP.v2_annotation (annotation_id, pgrm_gene_num,
template_id, strand, feature_type, feature_start, feature_stop, feature_length, exon_num,
program_id, update_date ) values ($aid, $Gene, \'$tid\', \'$Strand\', \'$Type\', $ExonB,  $ExonE,
$ExonL, $Exon, $pid, \'$update_date\')");
$sth->execute();
$sth->finish();

$dbh->disconnect;
}
}
close GSCANOUT;
}
```

Appendix E: Perl Script Genscan2Ora.pl. This script was designed and implemented by A. Eastman and D. Kolychev.

```
#!/usr/bin/perl -w

#use strict;
use LWP::Simple;
use lib ("/home/peela");
use uga_perl::util::util_oradbi
       qw(getConnection);
my @filearr;
my %maxexon;
my %maxexonf;

#------------------------------ Main ------------------------------------------------

## ---- Check Command Line Argument ----

if (@ARGV != 4) {
  die " Invalid Arguments! \n Usage: <scriptname>
<template_id><genemark_output><program_id><update_date>\n";
}

Insert2Ora($ARGV[0],$ARGV[1],$ARGV[2],$ARGV[3]);

#------------------------------ End of Main ----------------------------------------


#------------------------------ Insert2Ora ----------------------------------------

sub Insert2Ora {

my $readline="";
my($tid,$GScanOutFile,$pid,$update_date) = @_ ;

open (GSCANOUT, $GScanOutFile) or die "Cannot open the input file: $GScanOutFile \n";

while ($readline = <GSCANOUT>) {

   chomp($readline);
   push (@filearr, $readline);
   my @GeneExon;
   my $Gene;
   my $Exon;
   my $UniqueID;
   my $Type;
```

```
my $Strand;
my $ExonB;
my $ExonE;
my $ExonL;
my $Fr;
my $Ph;
my $IAc;
my $DoT;
my $CodRg;
my $P;
my $Tscr;

if ($readline =~ /Term|Init|Intr|PlyA|Prom/) {

        my @fields = split (' ',$readline);

    @GeneExon = split('\.',$fields[0]);
    $Gene   = $GeneExon[0];
    $Exon   = $GeneExon[1];
    $UniqueID = "$Gene.$Exon";


  if ($readline =~ /Term|Init|Intr/) {

    $Type   = $fields[1];
    $Strand = $fields[2];
    $ExonB  = $fields[3];
    $ExonE  = $fields[4];
    $ExonL  = $fields[5];
    $Fr     = $fields[6];
    $Ph     = $fields[7];
    $IAc    = $fields[8];
    $DoT    = $fields[9];
    $CodRg  = $fields[10];
    $P      = $fields[11];
    $Tscr   = $fields[12];
  }
  else {

    $Type   = $fields[1];
    $Strand = $fields[2];
    $ExonB  = $fields[3];
    $ExonE  = $fields[4];
    $ExonL  = $fields[5];
    $Fr     = "NULL";
    $Ph     = "NULL";
```

```
        $IAc    = "NULL";
        $DoT    = "NULL";
        $CodRg  = "NULL";
        $P      = "NULL";
        $Tscr   = $fields[6];
      }

      if ($Strand eq "+") {
            $Strand = "plus";
      }
      elsif ($Strand eq "-") {
            $Strand = "minus";
      }

      if($Type eq "Init"){
            .....$Type="Initial";
      }
      elsif($Type eq "Term"){
            .....$Type="Terminal";
      }
      elsif($Type eq "Intr"){
            .....  $Type="Internal";
      }
      elsif($Type eq "PlyA"){
            .....  $Exon="NULL";
      }

if($readline =~ /Term|Init|Intr/){
            if(!$maxexon->{$Gene}){
            ..... $maxexon->{$Gene}=1;
            .....$maxexonf->{$Gene}=1;
            }
            else{
            .....$maxexon->{$Gene}++;
            .....$maxexonf->{$Gene}++;
            }
}
}
}
close GSCANOUT;

foreach $readline(@filearr){
my @GeneExon;
   my $Gene;
   my $Exon;
   my $UniqueID;
```

```
my $Type;
my $Strand;
my $ExonB;
my $ExonE;
my $ExonL;
my $Fr;
my $Ph;
my $IAc;
my $DoT;
my $CodRg;
my $P;
my $Tscr;

if ($readline =~ /Term|Init|Intr|PlyA|Prom/) {

        my @fields = split (' ',$readline);

    @GeneExon = split('\.',$fields[0]);
    $Gene   = $GeneExon[0];
    $Exon   = $GeneExon[1];
    $UniqueID = "$Gene.$Exon";
  if ($readline =~ /Term|Init|Intr/) {

    $Type   = $fields[1];
    $Strand = $fields[2];
    $ExonB  = $fields[3];
    $ExonE  = $fields[4];
    $ExonL  = $fields[5];
    $Fr     = $fields[6];
    $Ph     = $fields[7];
    $IAc    = $fields[8];
    $DoT    = $fields[9];
    $CodRg  = $fields[10];
    $P      = $fields[11];
    $Tscr   = $fields[12];
  }
  else {

    $Type   = $fields[1];
    $Strand = $fields[2];
    $ExonB  = $fields[3];
    $ExonE  = $fields[4];
    $ExonL  = $fields[5];
    $Fr     = "NULL";
    $Ph     = "NULL";
    $IAc    = "NULL";
```

```
        $DoT    = "NULL";
        $CodRg  = "NULL";
        $P      = "NULL";
        $Tscr   = $fields[6];
    }
    if ($Strand eq "+") {
            $Strand = "plus";
    }
    elsif ($Strand eq "-") {
            $Strand = "minus";
    }

    if($Type eq "Init"){
        .....$Type="Initial";
    }
    elsif($Type eq "Term"){
        .....$Type="Terminal";
    }
    elsif($Type eq "Intr"){
        .....  $Type="Internal";
    }
    elsif($Type eq "PlyA" || $Type eq "Prom"){
        .....  $Exon="NULL";
    }
if($Strand eq "plus"){
my $DB_Type="test";
my $Oracle_ID="version2_GP";
my $Oracle_Passwd="version2_GP";
my $dbh = getConnection($DB_Type, $Oracle_ID, $Oracle_Passwd);

my $sth = $dbh->prepare("insert into version2_GP.v2_genscan (gene_num, template_id,
feature_start, frame, net_phase_exon, init_splice_score, donor_term_score, cod_reg_score,
probability, exon_score, program_id, update_date) values ($Gene, \'$tid\', $ExonB, $Fr, $Ph,
$IAc, $DoT, $CodRg, $P, $Tscr, $pid, \'$update_date\')");
$sth->execute();
$sth->finish();

$sth = $dbh->prepare('select annotation_id.nextval from dual');
$sth->execute();
my @ary = $sth->fetchrow_array();
my $aid=shift @ary;
$sth->finish();

if($readline =~ /Term|Init|Intr/){
        .....if($maxexonf->{$Gene}==1){
        ..... $Type="Single";
```

```
        ..... $Exon=1;
        ..... }
}

if($Type eq "PlyA" || $Type eq "Prom"){
        .....    $Exon="NULL";
}

$sth = $dbh->prepare("insert into version2_GP.v2_annotation (annotation_id, pgrm_gene_num,
template_id, strand, feature_type, feature_start, feature_stop, feature_length, exon_num,
program_id, update_date ) values ($aid, $Gene, \'$tid\', \'$Strand\', \'$Type\', $ExonB,  $ExonE,
$ExonL, $Exon, $pid, \'$update_date\')");
$sth->execute();
$sth->finish();

$dbh->disconnect;
}
elsif($Strand eq "minus"){
my $DB_Type="test";
my $Oracle_ID="version2_GP";
my $Oracle_Passwd="version2_GP";
my $dbh = getConnection($DB_Type, $Oracle_ID, $Oracle_Passwd);

my $sth = $dbh->prepare("insert into version2_GP.v2_genscan (gene_num, template_id,
feature_start, frame, net_phase_exon, init_splice_score, donor_term_score, cod_reg_score,
probability, exon_score, program_id, update_date) values ($Gene, \'$tid\', $ExonB, $Fr, $Ph,
$IAc, $DoT, $CodRg, $P, $Tscr, $pid, \'$update_date\')");
$sth->execute();
$sth->finish();

$sth = $dbh->prepare('select annotation_id.nextval from dual');
$sth->execute();
my @ary = $sth->fetchrow_array();
my $aid=shift @ary;
$sth->finish();
if($readline =~ /Term|Init|Intr/){
        .....$Exon=$maxexon->{$Gene};
        .....$maxexon->{$Gene}--;
}
if($readline =~ /Term|Init|Intr/){
        .....if($maxexonf->{$Gene}==1){
        ..... $Type="Single";
        ..... $Exon=1;
        ..... }
}
if($Type eq "PlyA" || $Type eq "Prom"){
```

..... $Exon="NULL";
}
$sth = $dbh->prepare("insert into version2_GP.v2_annotation (annotation_id, pgrm_gene_num, template_id, strand, feature_type, feature_start, feature_stop, feature_length, exon_num, program_id, update_date) values ($aid, $Gene, \'$tid\', \'$Strand\', \'$Type\', $ExonB, $ExonE, $ExonL, $Exon, $pid, \'$update_date\')");
$sth->execute();
$sth->finish();

$dbh->disconnect;
}
}
}
}

Appendix E: Perl Script RiceHMM2Ora.pl. This script was designed and implemented by A. Eastman and D. Kolychev.

```
#!/usr/bin/perl -w

#use strict;
use LWP::Simple;
use lib ("/home/peela");
use uga_perl::util::util_oradbi
       qw(getConnection);


my @filearr;
my %maxexonf;
my %maxexonb;

#----------------------------- Main ------------------------------------------------

## ---- Check Command Line Argument ----

if (@ARGV != 4) {
  die " Invalid Arguments! \n Usage: <scriptname>
<template_id><genemark_output><program_id><update_date>\n";
}


Insert2Ora($ARGV[0],$ARGV[1],$ARGV[2],$ARGV[3]);


#----------------------------- End of Main ----------------------------------------


#----------------------------- Insert2Ora ----------------------------------------



sub Insert2Ora {

my $readline="";
my($tid,$GScanOutFile,$pid,$update_date) = @_ ;

open (GSCANOUT, $GScanOutFile) or die "Cannot open the input file: $GScanOutFile \n";

while ($readline = <GSCANOUT>) {

   chomp($readline);
```

```
push (@filearr, $readline);
my @GeneExon;
my $Gene;
my $Exon;
my $UniqueID;
my $Type;
my $Strand;
my $ExonB;
my $ExonE;
my $ExonL;
my $Fr;
my $Ph;
my $IAc;
my $DoT;
my $CodRg;
my $P;
my $Tscr;
my $ORFB;
my $ORFE;
my $Score;
my $Index;
my $Frame;

if ($readline =~ /3'UTR|Terminal|Initial|Internal|Single/) {

        my @fields = split (' ',$readline);

    $Gene   = $fields[0];
    $Strand = $fields[1];
    $Type   = $fields[2];
    $ExonB  = $fields[3];
    $ExonE  = $fields[4];
    $ExonL  = $fields[5];
    $Index  = $fields[6];
    $Frame  = $fields[7];
        $Exon="NULL";
  if($Frame eq "-"){
      ..... $Frame="NULL";
  }
  if($Type=~/3/){
      .....$Type="3primeUTR";
  }

  if ($Strand eq "+") {
        $Strand = "plus";
  }
```

```
        elsif ($Strand eq "-") {
                $Strand = "minus";
        }

if($Strand eq "plus"){
my $DB_Type="test";
my $Oracle_ID="version2_GP";
my $Oracle_Passwd="version2_GP";
my $dbh = getConnection($DB_Type, $Oracle_ID, $Oracle_Passwd);

my $sth = $dbh->prepare("insert into version2_GP.v2_ricehmm (gene_num, template_id,
feature_start, score, frame, program_id, update_date) values ($Gene, \'$tid\', $ExonB, $Index,
$Frame, $pid, \'$update_date\')");
$sth->execute();
$sth->finish();

$sth = $dbh->prepare('select annotation_id.nextval from dual');
$sth->execute();
my @ary = $sth->fetchrow_array();
my $aid=shift @ary;
$sth->finish();

if($Strand eq "plus" && $readline =~ /Terminal|Initial|Internal|Single/){
                if(!$maxexonf->{$Gene}){
                ..........$maxexonf->{$Gene}=1;
                ..........$Exon=1;
                }
                else{
                ..........$maxexonf->{$Gene}++;
                ..........$Exon=$maxexonf->{$Gene};
                }
}


$sth = $dbh->prepare("insert into version2_GP.v2_annotation (annotation_id, pgrm_gene_num,
template_id, strand, feature_type, feature_start, feature_stop, feature_length, program_id,
update_date, exon_num) values ($aid, $Gene, \'$tid\', \'$Strand\', \'$Type\', $ExonB, $ExonE,
$ExonL, $pid, \'$update_date\', $Exon)");
$sth->execute();
$sth->finish();

$dbh->disconnect;
}
elsif($Strand eq "minus" && $readline =~ /Terminal|Initial|Internal|Single/){
                if(!$maxexonb->{$Gene}){
                .......... $maxexonb->{$Gene}=1;
```

```
            }
            else{
            ..... $maxexonb->{$Gene}++;
            }
}
}
}
close GSCANOUT;
foreach $readline(@filearr){
   my @GeneExon;
   my $Gene;
   my $Exon;
   my $UniqueID;
   my $Type;
   my $Strand;
   my $ExonB;
   my $ExonE;
   my $ExonL;
   my $Fr;
   my $Ph;
   my $IAc;
   my $DoT;
   my $CodRg;
   my $P;
   my $Tscr;
   my $ORFB;
   my $ORFE;
   my $Score;
   my $Index;
   my $Frame;

   if ($readline =~ /3'UTR|Terminal|Initial|Internal|Single/) {

          my @fields = split (' ',$readline);

      $Gene   = $fields[0];
      $Strand = $fields[1];
      $Type   = $fields[2];
      $ExonB  = $fields[3];
      $ExonE  = $fields[4];
      $ExonL  = $fields[5];
      $Index  = $fields[6];
      $Frame  = $fields[7];
          $Exon="NULL";
    if($Frame eq "-"){
          ..... $Frame="NULL";
```

```
        }
        if($Type=~/3/){
                .....$Type="3primeUTR";
        }

        if ($Strand eq "+") {
                $Strand = "plus";
        }
        elsif ($Strand eq "-") {
                $Strand = "minus";
        }

if($Strand eq "minus"){
my $DB_Type="test";
my $Oracle_ID="version2_GP";
my $Oracle_Passwd="version2_GP";
my $dbh = getConnection($DB_Type, $Oracle_ID, $Oracle_Passwd);

my $sth = $dbh->prepare("insert into version2_GP.v2_ricehmm (gene_num, template_id,
feature_start, score, frame, program_id, update_date) values ($Gene, \'$tid\', $ExonB, $Index,
$Frame, $pid, \'$update_date\')");
$sth->execute();
$sth->finish();

$sth = $dbh->prepare('select annotation_id.nextval from dual');
$sth->execute();
my @ary = $sth->fetchrow_array();
my $aid=shift @ary;
$sth->finish();

if($readline =~ /Terminal|Initial|Internal|Single/){
                .....$Exon=$maxexonb->{$Gene};
              .....$maxexonb->{$Gene}--;
}

$sth = $dbh->prepare("insert into version2_GP.v2_annotation (annotation_id, pgrm_gene_num,
template_id, strand, feature_type, feature_start, feature_stop, feature_length, program_id,
update_date, exon_num) values ($aid, $Gene, \'$tid\', \'$Strand\', \'$Type\', $ExonB, $ExonE,
$ExonL, $pid, \'$update_date\', $Exon)");
$sth->execute();
$sth->finish();
$dbh->disconnect;
}
}
}
}
```

Appendix G: Perl Script getart.pl. This script was designed and implemented by A. Eastman and D. Kolychev.

```
 #!/usr/bin/perl -w
use strict;
use LWP::Simple;
use lib ("/home/peela");
use uga_perl::util::util_oradbi
   qw(getConnection doSelect doNonSelect);
#----------------------------- Main -------------------------------------------------

  if (@ARGV != 4) {
  die " Invalid Arguments! \n Usage: <predictionFileName> <artemis_output_file>\n";
}
GetFromOra($ARGV[0],$ARGV[1]);

sub GetFromOra{
   my($tid, $artout) = @_ ;
   my @fieldarg=split("_", $tid);
   $tid=$fieldarg[0]."_".$fieldarg[1]."_".$fieldarg[2];
   my $pid=$fieldarg[3];
   open (OUT, ">$artout") or die "Cannot open the output file: $artout \n";
   my $DB_Type="test";
   my $Oracle_ID="version2_GP";
   my $Oracle_Passwd="version2_GP";
   my $dbh = getConnection($DB_Type, $Oracle_ID, $Oracle_Passwd);
   my $sth = $dbh->prepare("select max(PGRM_GENE_NUM) from v2_annotation where
template_id=\'$tid\' and program_id=$pid");
   $sth->execute();
   my @maxarr=$sth->fetchrow_array();
   my $max=shift @maxarr;
   my $c=0;
   my $strand;
   my $type;
   while($c<=$max){
          my $geneline="";
          $sth = $dbh->prepare("select feature_start, feature_stop, strand, feature_type from
v2_annotation where template_id=\'$tid\' and program_id=$pid and pgrm_gene_num=$c and
(feature_type='Initial' or feature_type='Internal' or feature_type='Terminal' or
feature_type='Single') order by feature_start");
          $sth->execute();
          my @genes;

          while(@genes = $sth->fetchrow_array()){
             #now join in the file if exists
             if(scalar(@genes)>0){
```

```
            #print "$genes[0] and $genes[1] at gene_num $c and strand $genes[2]\n";
          ..... $strand=$genes[2];
          ..... $type=$genes[3];
            if($geneline){
          ..... $geneline=$geneline.",$genes[0]..$genes[1]";
            }
            else{
           $geneline="$genes[0]..$genes[1]";
            }
          }
        }
         if($geneline){
           print "$geneline at genenum $c and strand $strand\n";
           my $pname="";
           $sth = $dbh->prepare("select program_name from v2_programs where
program_id=$pid");
           $sth->execute();
        my @namearr=$sth->fetchrow_array();
        $pname=shift @namearr;
           if($type ne "Single"){
          ..... $geneline="join(".$geneline.")";
           }

           if($strand eq "plus"){
          .....print OUT "FT   CDS             $geneline\n";
           }
           else{
          .....print OUT "FT   CDS             complement($geneline)\n";
           }
           print OUT "FT                   /note=\"predicted by $pname\"\n";
        print OUT "FT                   /codon_start=1\n";
           print OUT "FT                   /product=\"$pname.$c\"\n";
         }
         $sth->finish();
         $c++;
  }

  $dbh->disconnect;
}
```

Appendix H: Perl Script header.pl, This script was designed and implemented by A. Eastman and D. Kolychev.

```
#!/usr/bin/perl -w

#use strict;
use LWP::Simple;
use lib ("/home/peela");
use uga_perl::util::util_oradbi
       qw(getConnection);

#use uga_perl::utils::DBConnection
#    qw(getConnection doSelect doNonSelect);

#------------------------------ Main ------------------------------------------------

## ---- Check Command Line Argument ----

if (@ARGV != 2) {
  die " Invalid Arguments! \n Usage: <input_dir> <output_filename>\n";
}

my $indir=$ARGV[0];
my $outname=$ARGV[1];

my $parselist=`ls $indir`;
#print "got $parselist\n";
my @parr=split(" ", $parselist);

my %pghash;
$pghash->{genemark}="GM";
$pghash->{genscan}="GS";
$pghash->{ricehmm}="RH";
$pghash->{fgenesh}="FG";

foreach $f(@parr){
   print "processing $f\n";
   open (IN, "$indir/$f") or die ("cannot open $indir/$f\n");
   open (OUT, ">>$outname") or die ("cannot open $outname\n");
   my $name;
   my $ace;
   my $ab;
   my $des;
   my $start;
   my $stop;
   my $strand;
```

```
while(<IN>){
        if($_=~/^>/){
          my @params=split("_",$f);
          my @larr=split(" ",$_);
          $params[1]=~s/ace//;
          $ace=$params[1];
          $name=$params[0];
          if($f=~/_ORF_/){
          ..........$ab=$params[2];
          .......... $des=$params[4];
          ..........my @ss=split(":", $larr[2]);
          ..........if($des eq "nt"){
          ..... my @ss=split(":", $larr[2]);
          ..... $start=$ss[0];
          ..... $stop=$ss[1];
          ..... $strand=$larr[3];
          ..... }
          ..........elsif($des eq "aa"){
          ..... my @ss=split(":", $larr[3]);
          ..... $start=$ss[0];
          ..... $stop=$ss[1];
          ..... $strand=$larr[4];
          ..... }
          ..........if ($strand eq "forward"){
          ..... $strand="plus";
          ..... }
          ..... elsif($strand eq "reverse"){
          ..... $strand="minus";
          ..... }
          }
          elsif($f=~/_OVL_/){
          ..........$ab=$params[2];
          .......... $des=$params[3];
          ..........$_=~s/>//;
          ..... my @ss=split("_", $_);
          ..... $start=$ss[0]*100;
          ..... $stop=$ss[1]*100;
          ..... $strand="plus";
          .....
          }
          elsif($f=~/_MANseq_/){
          ..........$ab=$params[2];
          .......... $des=$params[3];
          ..........$start=$params[4];
          ..........$stop=$params[5];
          ..........$strand=$params[6];
```

```
.....if ($strand eq "forward"){
.....  $strand="plus";
..... }
..... elsif($strand eq "reverse"){
.....  $strand="minus";
..... }
  }
  else{
..... $ab=$pghash->{$params[2]};
..... $des=$params[3];
.....my @ss=split(":", $larr[2]);
..... $start=$ss[0];
..... $stop=$ss[1];
.....$strand=$larr[3];
.....if ($strand eq "forward"){
.....  $strand="plus";
..... }
..... elsif($strand eq "reverse"){
.....  $strand="minus";
..... }
  }

  my $newline=">".$name."|".$ace."|".$ab."|".$des."|".$start."|".$stop."|".$strand."\n";

 if(!$name || !$ace || !$ab || !$des || !$strand){
  .....print "ERROR: something in $newline is undefined\n";
  ..... exit();
  }
  print OUT $newline;
 }
 else{
  print OUT $_;
 }
 }
}
```

Appendix I: Perl Script pf.pl. This script was designed and implemented by A. Eastman and D. Kolychev.

```
#!/usr/bin/perl -w

$inputfile = $ARGV[0] or die "Usage: <scriptname> <inputfile> <outputfile>\n";
$outputfile = $ARGV[1] or die "Usage: <scriptname> <inputfile> <outputfile>\n";

open (INFILE, $inputfile) or die "Cannot open $inputfile\n";
open (OUTFILE, ">$outputfile") or die "Cannot open $outputfile\n";

$start=0;
$end=30;


  while (<INFILE>) {
        $readline = $_;
        if($readline=~ m/>/){
           $readline='>'.$start.'_'.$end;
           $start=$start+15;
           $end=$end+15;
           print OUTFILE ("$readline\n");

        }
        else{
           print OUTFILE ("$readline");
        }
   }


close INFILE;
close OUTFILE;
```

Appendix J: Perl Script BlastToMagic.pl. This script was designed, written, and implemented by M.-M. Cordonnier-Pratt and F. Sun.

```
#!/usr/local/bin/perl -w

eval 'exec /usr/local/bin/perl -w -S $0 ${1+"$@"}'
   if 0; # not running under some shell
eval 'exec /usr/bin/perl -w -S $0 ${1+"$@"}'
   if 0; # not running under some shell

##############################################################################
# Copyright @ 2003, Laboratory for Genomics and Bioinformatics LGB          #
# at the University of Georgia                                              #
# All rights reserved.                                                      #
#                                                                           #
# This software is provided "AS IS".  UGA-LGB makes no warranties, express  #
# or implied, including no representation or warranty with respect to       #
# the performance of the software and derivatives or their safety,          #
# effectiveness, or commercial viability.  UGA-LGB does not warrant the     #
# merchantability or fitness of the software and derivatives for any        #
# particular purpose, or that they may be exploited without infringing      #
# the copyrights, patent rights or property rights of others. UGA and LGB   #
# shall not be liable for any claim, demand or action for any loss, harm,   #
# illness or other damage or injury arising from access to or use of the    #
# software or associated information, including without limitation any      #
# direct, indirect, incidental, exemplary, special or consequential         #
# damages.                                                                  #
#                                                                           #
# This software program may not be sold, leased, transferred, exported      #
# or otherwise disclaimed to anyone, in whole or in part, without the       #
# prior written consent of LGB-UGA                                          #
##############################################################################
#
# Change History
# 031002, fsun, add hsp_tot_num for each hit
#
#************************************************************
# 1. Documentation Summary
#************************************************************
=head1 NAME
blastToMagic_main.pl
=head1 SYNOPSIS
   blastToMagic_main.pl config_file

=head1 DESCRIPTION
  MAGIC Database & Software is a highly Oracle DBMS-integrated system.
```

```
  QVS sequence stands for the sequence with quality 16, no vector and
  no contamination from E. coli and ribosomal RNA.
  magic_qvsbase is used to process QVS sequences and determine the
  number and proportion of each base (A,T,G,C and N). Also,
  magic_qvsbase will help to determine sequence with multi-A,-T,-G or
  -C as well as G-C ratio for each sequence.
  blastToMagic_main.pl - program to parse Blast output using BioPerl and
  to insert the data into Magic database

=head1 Mailing List
=head2 Reporting Bugs
=head1 AUTHOR - Chun Liang, Feng Sun
Email cliang@uga.edu fsun@uga.edu
=cut

use strict;
use Bio::SearchIO;
use File::Basename;
use Getopt::Long;
use DBI;
use Config::IniFiles;

use lib ("../lib");
#use lib ("/usr/local/lib");
use uga_perl::util::util_oradbi
   qw(getConnection
      beginTransaction
      endTransaction
      abortTransaction);
use uga_perl::util::util_systool
   qw(DirOrFileExist
      ExeSysCmd);
use uga_perl::util::util_time
   qw(GetTimeInNum GetDate);
use uga_perl::util::MakeFastaQual
   qw(WriteFastaFileFromDB);
use uga_perl::bio::bio_ace
   qw( CreateAnalysisGroup
       GetAnalysisProgramId
       TgiclToStdAce
       countAce
       parseAceToDb
       toClass012Ace
       singletonFileToAceFile);
```

```
##############################################################################
# start  Main
##############################################################################

#global variables
$::Date_Last_Modified = "Apr. 21st, 2004";
($::my_name) = $0 =~ m"[\\/]?([^\\/]+)$";
$::my_name ||= 'blastToMagic_main.pl';

## get command-line arguments, or die with a usage statement
my $usage = qq{
blastToMagic_main.pl - Program to parse Blast output into Magic database
  version 0.93.2, Feng Sun, $::Date_Last_Modified
Usage:
  blastToMagic_main.pl configfile blast_file1 [blast_file2 ...]
where:
  blast_files are output files from Blast
  A sample configuration file:

####################### blastToMagic_main.pl ###################################
[database]
db_name=demodata
oracle_id=magic
oracle_passwd=magic

[blastToMagic]
## this section contains parameters for blastToMagic_main.pl, which parses
#       result files from Blast to MAGIC database system

## database id of query sequences, defined in table v2_blast_database
# Two default database ids will be created when installing MAGIC system
# 1, database id for est sequences in MAGIC database, which is the default query
#    database id for Blast
# 2, database id for PIR-NREF protein sequences, which is the default target
#    database id for Blast
query_database_id=1
## database gric is used to differentiate different releases of a same database
# It starts from 1 and increment
query_database_gric=1
## quality process id is a field associated with each processed EST sequences which
# is used to record the methods for processing trace files. If a database doesn't
# has this feature, use default value 1
query_quality_process_id=1
## database id of query sequences, defined in table v2_blast_database
# Two default database ids will be created when installing MAGIC system
# 1, database id for est sequences in MAGIC database, which is the default query
```

```
#    database id for Blast
# 2, database id for PIR-NREF protein sequences, which is the default target
#    database id for Blast
target_database_id=2
## database gric is used to differentiate different releases of a same database
# It starts from 1 and increment
target_database_gric=1
## quality process id is a field associated with each processed EST sequences which
# is used to record the methods for processing trace files. If a database doesn't
# has this feature, use default value 1
target_quality_process_id=1
## set debug mode, 0 or 1
debug_on=0
};

$::ConfigFileName  = $ARGV[0] or die $usage;
my $cfg = new Config::IniFiles( -file => $::ConfigFileName );
####variables
$::DB_Name         = $cfg->val('database', 'db_name');
$::Oracle_ID       = $cfg->val('database', 'oracle_id');
$::Oracle_Passwd   = $cfg->val('database', 'oracle_passwd');

my $targetDbId  = $cfg->val('blastToMagic', 'target_database_id');
$::DBID = $targetDbId;   ##old name for targetDbId
my $targetDbGric  = $cfg->val('blastToMagic', 'target_database_gric');
$::GRIC = $targetDbGric;   ##old name for targetDbId
my $targetQualId  = $cfg->val('blastToMagic', 'target_quality_process_id');

my $queryDbId  = $cfg->val('blastToMagic', 'query_database_id');
my $queryDbGric  = $cfg->val('blastToMagic', 'query_database_gric');
my $queryQualId  = $cfg->val('blastToMagic', 'query_quality_process_id');
$::DEBUG = $cfg->val('blastToMagic', 'debug_on');  #0 or 1

##global variable $processDate, used in v2_blast_run and v2_database_sequence
my $processDate = GetDate();

if ( $::DEBUG eq '0' )
{
  $::DEBUG = '';
}
else
{
  $::DEBUG = 1;
}

print "debug: $::DEBUG\n";
```

```
$::BLAST_RUN_ID = '';

$::total_hsps = $::total_queries = $::total_files = 0;
my @hsp_list = ();
#shift out config_file from argument list
shift @ARGV;
die 'You must provide a file name' unless @ARGV;

# process list of file names from command line
processBlastFiles(@ARGV);

################################################################################
# End Main
################################################################################

################################################################################
# process blast object from query
################################################################################

sub processBlastFiles
{
   my (@bfs) = @_;

   ############--get connection
   my $dbh = getConnection($::DB_Name, $::Oracle_ID, $::Oracle_Passwd);
   print "\n\tdatabase connection created \n" if $::DEBUG;
   beginTransaction($dbh);

   if ($::DEBUG){
         print "input files: @bfs\n";
   }

   #process each files
   foreach my $file (@bfs)
   {
      my ($searchin, $query);
      my $file_queries = 0;
      $::total_files++;
      print "\nprocessing file $::total_files $file \n" ;
      $searchin = new Bio::SearchIO( -format => 'blast', -file => $file );
      # $searchin should be a Bio::SearchIO object
      unless ($searchin)
      {
         print STDERR "Call to new Bio::SearchIO failed for file: '$file'\n";
         next;
      }
```

```
      print "\nnew Bio::SearchIO created for file  $file \n" if $::DEBUG;

      #--create one runId for input files
      my $updateRun = 0;

      #process this file
      #for each file, we will create a run id

      while ( $query = $searchin->next_result() ) # next query within output file
      {
         # $query is a Bio::Search::Result::ResultI object
         $file_queries++;
            if ($updateRun == 0){
               #create a blast_run_id for this file
               updateBlastRun($dbh, $query, $fileName);
            $updateRun =  1;
         }
            else{
            processBlast($dbh, $query);
         }
      }
         ## update qual_process_id and pir_nref in v2_database_sequence
         updateQualProcessId($dbh);
      print STDERR "  $file_queries blast queries parsed from file: $file\n";
         $::total_queries += $file_queries;
   } # end foreach $file (@ARGV)

   print STDERR "\n$::total_files files containing $::total_queries queries were found with
$::total_hsps HSPs\n";
   #--disconnect from database
   endTransaction($dbh);
   $dbh->disconnect;
   print "\n\tdatabase connection droped \n" if $::DEBUG;
}
sub updateBlastRun
{
   my ($dbh, $query, $blastFileName) = @_;
   #--stuff to update blast run info

   my ($inputFileName, $inputDesc, $viewFileName);

   my $blastType = $query->algorithm();
   my $matrix = $query->get_parameter('matrix');
   #--get list of parameters
   my @params = $query->available_parameters();
   my $strParams = '';
```

```
    my $i = 0;
    for ( $i = 0; $i <= $#params; $i++ ){
          if ( $params[$i] ne 'matrix' ){
          $strParams .= $params[$i] . "="
                        . $query->get_parameter($params[$i]) . " ";
       }
    }
    print "\nupdating blast run info....\n" if $::DEBUG;
    #--test connection
    #testConnection($DB_Type, $Oracle_ID, $Oracle_Passwd);

    my $processTime = GetTimeInNum();
    #--update run table
    my $sqlRun = qq{
          insert into v2_blast_run (blast_run_id, process_date,
               process_time, input_file_name, input_description,
             blast_output_file_name, view_file_name, query_database_id,
             query_database_gric, target_database_id, target_database_gric,
                blast_type, matrix, parameters)
             values( SEQ_102_BLAST_RUN_ID.nextval,?,?,?,?,?,?,?,?,?,?,?,?,?)
    };

    my $sth = $dbh->prepare( $sqlRun )
       or die "$::my_name :UpdateBlastRun:Couldn't prepare statement:".$dbh->errstr;

    print qq{ $sqlRun: date: $processDate, time: $processTime, blastFile: $blastFileName,
                databaseId: $::DBID, gric: $::GRIC, algm.: $blastType, matrix: $matrix,
              parameters: $strParams \n} if $::DEBUG;
    $sth->execute($processDate, $processTime, $inputFileName, $inputDesc,
          $blastFileName,$viewFileName, $queryDbId, $queryDbGric, $::DBID, $::GRIC,
$blastType, $matrix,
          $strParams)
       or die "Blast2V2:UpdateBlastRun:Couldn't execute statement:".$sth->errstr;
    print "finished updating blast run info....\n" if $::DEBUG;
    $::BLAST_RUN_ID = getRunIdByTime($dbh, $processDate, $processTime);
    print "\nBLAST_RUN_ID: $::BLAST_RUN_ID\n" if $::DEBUG;
    #--add this result to database
    processBlast($dbh, $query);
}
sub updateQualProcessId{
    my ($dbh) = @_;
    my $strSql = '';

    if ( $queryDbId == 1 ){ ##EST sequence
      $strSql = qq{
          update v2_database_sequence A
```

```
        set ( A.qual_process_id, A.magic_seq_id) =
        (select qual_process_id, magic_seq_id
         from v2_seq_qual_stats B
         --For Harvard sequences, alias_name is saved in Blast table following instruction
         --from Marie-michele. This (The mixing of names) is a potential problem.
         where ( A.sequence_name = B.seq_name or A.sequence_name = B.alias_name)
              and B.tag_process_activity = '1'
        )
        --1 is the est database
        where A.database_id = '1' and
             A.SEQ_UPDATE_DATE = '$processDate'
    };
  }
  elsif ( $queryDbId == 4 ){ ##UniScript sequence
    $strSql = qq{
      update v2_database_sequence A
      set ( A.magic_seq_id) =
      (select B.ctg_magic_seq_id
       from v2_analysis_cluster_ctg B
       where A.sequence_name = B.contig_id and A.qual_process_id = B.analysis_run_id
      )
      --4 is the analysis cluster
      where A.database_id = '4' and
           A.SEQ_UPDATE_DATE = '$processDate'
    };
  }
  if ( $queryDbId == 1 or $queryDbId == 4){
    print "Updating MAGIC seq id: $strSql\n";
    my $numRows = $dbh->do($strSql);
    print "Rows changed: $numRows\n";
  }
  #die "Error: no rows updated for qual_process_id\n" if $numRows <= 0;
}
sub testConnection{
  my($DB_Type, $Oracle_ID, $Oracle_Passwd) = @_;
  print "testting database connection.....\n";
  my $dbh = getConnection($DB_Type, $Oracle_ID, $Oracle_Passwd);
  my $sth = $dbh->prepare('select sysdate from dual');
  $sth->execute();
  my @ary = $sth->fetchrow_array();
  $sth->finish();
  print "\nresult from oracle: @ary \n\n";
  $dbh->disconnect;
}

##############################################################################
```

```
# get blast run id by processDate and processTime
###############################################################################
sub getRunIdByTime{
   my($dbh, $processDate, $processTime) = @_;
   #my $dbh = getConnection($DB_Type, $Oracle_ID, $Oracle_Passwd);
   #DBI->trace( 2 );
   my $sth = $dbh->prepare('select BLAST_RUN_ID from V2_BLAST_RUN WHERE '.
         'PROCESS_DATE=? AND PROCESS_TIME=?');
   $sth->execute($processDate, $processTime)
      or die "$::my_name :getRunIdByTime:Couldnot execute statement:".$sth->errstr;
   my @ary = $sth->fetchrow_array();
   $sth->finish();
   print "date:$processDate, time:$processTime, blast_run_id:$ary[0]\n" if $::DEBUG;
   return $ary[0];
}

###############################################################################
# process blast object from query
# Steps:
#   1. Get query id in v2_database_sequence, if a sequence has been uploaded
#      before, return that id. Otherwise, create a new one.
#   2. Read hits for this query. Get target id for each target.
#   3. Get HSP for each hit and insert into v2_blast_hsp
###############################################################################

sub processBlast
{
   my ($dbh, $query) = @_; # $searchin is a Bio::Search::Result::ResultI object

   my $hit;
   my (@temp_hsp_list, @sorted_hsps);
   my $hsps = 0;
   my $program  = $query->algorithm();       # BLASTN, BLASTP, ...
   my $DB       = $query->database_name();
   my $queryName = $query->query_name();
   my $queryAccession = $query->query_accession();
                              #bioPerl will return the same as query_name
   my $queryLen  = $query->query_length();
   my $queryDesc = $query->query_description();

   #--030803, FSUN, get sequenceId and updateDate together
   my $seqUpdateDate = GetDate();
   #--get query id, if it is the first in table database_sequence, insert it
   my ($queryId, $queryUpdateDate) = getDbSeqId($dbh, $seqUpdateDate, $queryAccession,
$queryDbId,
         $queryDbGric, $queryName, $queryQualId, $queryLen, $queryDesc);
```

```
   print "\n>$queryName, $queryId, $queryAccession, $queryLen \n" if $::DEBUG;
   print "processing query....\n" if $::DEBUG;

   while ($hit = $query->next_hit()) # returns Bio::Search::Hit::HitI objects
   {
      my $hitName    = $hit->name();
      my $hitLen     = $hit->length();
      my $hitAccession  = $hit->accession();
      my $hitDesc = $hit->description();
      ## 031002, fsun, add "hsp_tot_num"
      my $hspTotNum = $hit->num_hsps();
         #my $fakeQualId = 1; #fake id for hit, to make database consistent
      my $fakeQualId = $targetQualId; #fake id for hit, to make database consistent

      #--get query id, if it is the first in table database_sequence, insert it
      my ($hitId, $hitUpdateDate) = getDbSeqId($dbh, $seqUpdateDate, $hitAccession,
$::DBID,
            $::GRIC, $hitName, $fakeQualId, $hitLen, $hitDesc);

      print "->$hitName, $hitId, $hitAccession, $hitLen \n" if $::DEBUG;

      #--populate hsp table
      my $hitScore  = $hit->raw_score();
      my $hitEvalue = $hit->significance();
      $hitEvalue =~ s/^([eE][-+]?\d+)$/1$1/;             # 'E-123' => '1E-123'

      my $thsp;
         @temp_hsp_list = ();
      while ($thsp = $hit->next_hsp()) # $hsp is a Bio::Search::HSP::HSPI object
      {
            $hsps++;
         my $hspQueryStart = $thsp->start('query');
         my $hspQueryEnd = $thsp->end('query');

         push @temp_hsp_list,
            [$hspQueryStart, $hspQueryEnd, $thsp];
      } # end while ($hsp = ... )
      $thsp = undef;

         #--sort
      if ( scalar @temp_hsp_list >= 2)   #sort when at least two hsps
      {
         @sorted_hsps = sort by_pos @temp_hsp_list;
         @temp_hsp_list = @sorted_hsps;
            print "scalar @temp_hsp_list hsp were sorted \n" if $::DEBUG;
      }
```

```
            #--insert into BLAST_HSP
            print "insert HSP\n" if $::DEBUG;
         my $hspNum = 0;
            foreach my $savedHsp (@temp_hsp_list)
            {
               $hspNum++;
               my ($hspQueryStart, $hspQueryEnd, $hsp) = @{ $savedHsp };
               print "processing HSP: $hsp, $hspQueryStart, $hspQueryEnd\n" if $::DEBUG;
               insertBlastHsp($dbh, $hspNum, $hsp, $::BLAST_RUN_ID, $queryId,
                     $hitId, $hitScore, $hitEvalue, $hspTotNum);
            } # end for each
      } # end while ($hit =  ... )

      print STDERR "  For query=$queryName, $hsps HSPs were found\n"
        if ($::DEBUG);

      $::total_hsps += $hsps;
      $query = $hit = undef;
   } # end process_blast


   sub by_pos
   {
      my( $seq1bega, $seq1enda, $hspa) = @{ $a };
      my( $seq1begb, $seq1endb, $hspb) = @{ $b };
         $seq1bega <=> $seq1begb or     #   then beginning base position
         $seq1enda <=> $seq1endb;       #   then ending base position
   } # end by_pos
   sub getDbSeqId{
      my($dbh, $seqUpdateDate, $queryAccession, $queryDbId,
         $queryDbGric, $queryName, $queryQualId, $queryLen, $queryDesc) = @_;

      #--try seqs in other date
      my $seqId2 = '';
      my $seqDate = '';
      ($seqId2, $seqDate) = getDbSeqId2($dbh, $seqUpdateDate, $queryAccession, $queryDbId,
         $queryDbGric, $queryName, $queryQualId, $queryLen, $queryDesc);

      if ( not defined $seqId2 or $seqId2 eq ''){
         #--find totally new sequence, insert it to database
         insertDbSeqId1($dbh, $seqUpdateDate, $queryAccession, $queryDbId,
            $queryDbGric, $queryName, $queryQualId, $queryLen, $queryDesc);
         #--ge id for this new sequence
         ($seqId2, $seqUpdateDate) = getDbSeqId2($dbh, $seqUpdateDate, $queryAccession,
   $queryDbId,
         $queryDbGric, $queryName, $queryQualId, $queryLen, $queryDesc);
```

```
            print "new sequence, create id3: $seqId2 date:$seqUpdateDate for $queryAccession, "
                    . " $queryName\n" if $::DEBUG;
        ##--new sequence has today's date
        return ($seqId2, $seqUpdateDate);
    }
    else{
        #--sequence has been inserted on different date, use same i
        return ($seqId2, $seqDate);
    }
}
sub getDbSeqId1{
    my($dbh, $seqUpdateDate, $queryAccession, $queryDbId,
        $queryDbGric, $queryName, $queryQualId, $queryLen, $queryDesc) = @_;

   ## no need to compare qual_process_id because there is no way to know it
    my $sth1 = $dbh->prepare(
            qq{select distinct database_sequence_id
        from v2_database_sequence where db_accession_num=? and database_id=?
        and gric=? and sequence_name=?
            and seq_update_date=? } );
        ##and gric=? and sequence_name=? and qual_process_id=?
    $sth1->execute($queryAccession, $queryDbId, $queryDbGric, $queryName,
        $seqUpdateDate);
        #$queryQualId, $seqUpdateDate);
    my @ary;
    #--should get only one row
    while( @ary = $sth1->fetchrow_array()){
        my $dbSeqId = $ary[0];
        return ($dbSeqId, $seqUpdateDate);
    }
    return ('', '');
}
sub getDbSeqId2{
    my($dbh, $seqUpdateDate, $queryAccession, $queryDbId,
        $queryDbGric, $queryName, $queryQualId, $queryLen, $queryDesc) = @_;

    #--if multiple entries have same seqId, return the oldest one
    my $sth1 = '';

    if ( $queryDbId eq '1' ){
        $sth1 = $dbh->prepare(
        qq { select distinct database_sequence_id, seq_update_date
         from v2_database_sequence where db_accession_num=? and database_id=?
         and gric=? and sequence_name=? order by seq_update_date});
        $sth1->execute($queryAccession, $queryDbId, $queryDbGric, $queryName);
    }
```

```
    else{
        $sth1 = $dbh->prepare(
        qq { select distinct database_sequence_id, seq_update_date
         from v2_database_sequence where db_accession_num=? and database_id=?
         and gric=? and sequence_name=? and qual_process_id=? order by seq_update_date});
        $sth1->execute($queryAccession, $queryDbId, $queryDbGric, $queryName,
$queryQualId);
    }

    my @ary;
    while( @ary = $sth1->fetchrow_array()){
        my $dbSeqId = $ary[0];
        my $dbSeqDate = $ary[1];
        #if ($seqUpdateDate eq $dbSeqDate){
            #$sth->finish();
        return ($dbSeqId, $dbSeqDate);
    }
    return ('', '');
}
sub insertDbSeqId1{
    my($dbh, $seqUpdateDate, $queryAccession, $queryDbId, $queryDbGric,
          $queryName, $queryQualId, $queryLen, $queryDesc) = @_;

    my $strSql = qq{
        insert into v2_database_sequence
        (database_sequence_id, seq_update_date, db_accession_num,
         database_id, gric, sequence_name, qual_process_id, sequence_length,
         sequence_description) values (seq_103_database_sequence_id.nextval,
         ?, ?, ?, ?, ?, ?, ?, ?)
    };
    my $sth1 = $dbh->prepare($strSql);
    print "------>New database sequence\n$strSql" if $::DEBUG;
    print qq{-->values\n$seqUpdateDate, $queryAccession, $queryDbId, $queryDbGric,
        $queryName, $queryQualId, $queryLen, $queryDesc} if $::DEBUG;
    $sth1->execute($seqUpdateDate, $queryAccession, $queryDbId, $queryDbGric,
          $queryName, $queryQualId, $queryLen, $queryDesc);

    #todo, populate pir table?
}
sub insertBlastHsp{
    my ($dbh, $hspNum, $hsp, $bRunId, $queryId, $hitId,
                    $hitScore, $hitEvalue, $hspTotNum) = @_;
    my $score  = $hsp->score();
    my $bits   = $hsp->bits();
    my $evalue = $hsp->evalue();
    $evalue =~ s/^([eE][-+]?\d+)$/1$1/; # 'E-123' => '1E-123'
```

```
my $hspStrand = $hsp->strand();
my $hspQueryFrame = $hsp->frame();
my $convertedFrame = ($hspQueryFrame + 1) * $hspStrand;
print "frame: $convertedFrame, ($hspStrand, $hspQueryFrame)\n" if $::DEBUG;
my $hspQueryStart = $hsp->start('query');
my $hspQueryEnd = $hsp->end('query');
my $hspHitStart = $hsp->start('hit');
my $hspHitEnd = $hsp->end('hit');
my $hspQSeq = $hsp->query_string();
my $hspQSeq2 = '';
if ( length($hspQSeq) > 4000 ){
        $hspQSeq2 = substr($hspQSeq, 4000);
        $hspQSeq = substr($hspQSeq, 0, 4000);
}
my $hspMidLine = $hsp->homology_string();
my $hspMidLine2 = '';
if ( length($hspMidLine) > 4000 ){
        $hspMidLine2 = substr($hspMidLine, 4000);
        $hspMidLine = substr($hspMidLine, 0, 4000);
}
my $hspHSeq = $hsp->hit_string();
my $hspHSeq2 = '';
if ( length($hspHSeq) > 4000 ){
        $hspHSeq2 = substr($hspHSeq, 4000);
        $hspHSeq = substr($hspHSeq, 0, 4000);
}
my @qInds =  $hsp->seq_inds('query', 'identical');
my $hspQueryInds = join(',', @qInds );
##print "query identical: @qInds \n" if $::DEBUG;
my $hspHitInds = join(',', $hsp->seq_inds('hit', 'identical') );
my $hspGaps = $hsp->gaps();
my $hspAlignLength = $hsp->length('total');
my $hspIdentity = $hsp->num_identical();
my $hspPositive = $hsp->num_conserved();
my $percentIdentity = $hsp->frac_identical * 100;
my $percentPositive = $hsp->frac_conserved * 100;
my $hspRank = $hsp->rank();
my $sth = $dbh->prepare(
        qq { insert into v2_blast_hsp
    (blast_run_id, query_id, target_id,
         hsp_num, SCORE, EXPECT, HSP_STRAND, HSP_QUERY_FRAME, HSP_SCORE,
         HSP_QUERY_START, HSP_QUERY_END, HSP_HIT_START, HSP_HIT_END,
         HSP_QSEQ, HSP_MIDLINE, HSP_HSEQ,
         HSP_QSEQ2, HSP_MIDLINE2, HSP_HSEQ2,
         HSP_GAPS, HSP_ALIGN_LENGTH, HSP_IDENTITY, HSP_POSITIVE,
```

```
          PERCENT_IDENTITY, PERCENT_POSITIVE, HSP_EXPECT, RANK,
HSP_TOT_NUM)
        values (?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?, ?,
                  ?,?,?,?,?,?,?,?,?,?,?) } );
  print "****new hsp to be inserted: ($bRunId, $queryId, $hitId,
         $hspNum, $hitScore, $hitEvalue, $hspStrand, $convertedFrame, $score, $hspQueryStart,
       $hspQueryEnd, $hspHitStart, $hspHitEnd,
         $hspQSeq, $hspMidLine, $hspHSeq,
         $hspQSeq2, $hspMidLine2, $hspHSeq2,
       $hspGaps, $hspAlignLength, $hspIdentity,
       $hspPositive, $percentIdentity, $percentPositive, $evalue, $hspRank, $hspTotNum) \n" if
$::DEBUG;
   $sth->execute($bRunId, $queryId, $hitId, $hspNum, $hitScore,
         $hitEvalue, $hspStrand, $convertedFrame, $score, $hspQueryStart,
         $hspQueryEnd, $hspHitStart, $hspHitEnd,
         $hspQSeq, $hspMidLine, $hspHSeq,
         $hspQSeq2, $hspMidLine2, $hspHSeq2,
         $hspGaps, $hspAlignLength, $hspIdentity,
         $hspPositive, $percentIdentity, $percentPositive, $evalue, $hspRank, $hspTotNum);
}
##############################################################################
# by_expect - sort comparison routine to sort by increasing expect value
#   or by decreasing score if expect values match
##############################################################################
sub by_expect
   {
   my($evaluea, $scorea) = @{ $a };
   my($evalueb, $scoreb) = @{ $b };
   $evaluea <=> $evalueb or $scoreb <=> $scorea;
   } # end by_expect
##############################################################################
# by_score - sort comparison routine to sort by decreasing score
#   or by increasing expect value if scores match
##############################################################################
sub by_score
   {
   my($evaluea, $scorea) = @{ $a };
   my($evalueb, $scoreb) = @{ $b };
   $scoreb <=> $scorea or $evaluea <=> $evalueb;
   } # end by_score
```

Appendix K: Perl Script artfromblast.pl. This script was designed and implemented by A. Eastman and D. Kolychev.

```
#!/usr/bin/perl -w

use strict;
use LWP::Simple;
use lib ("/home/peela");
use uga_perl::util::util_oradbi
    qw(getConnection doSelect doNonSelect);

#------------------------------ Main ------------------------------------------------

## ---- Check Command Line Argument ----

if (@ARGV != 1) {
   die " Invalid Arguments! \n Usage: <config filename>\n";
}


GetFromOra($ARGV[0]);

sub GetFromOra{
  my $config=shift @_;
  my $tid;
  my $artout;
  my $inputsql;
  my $variable;
  my $colour;
  open (IN, "<$config") or die "Cannot open the config file: $config \n";
  while(<IN>){
     $_=~s/\n//;
     my @param=split("=",$_);
     if ($param[0] eq "variable"){
            $variable=$param[1];
     }
     if ($param[0] eq "blastfile"){
            $tid=$param[1];
     }
     if ($param[0] eq "outfile"){
            $artout=$param[1];
     }
     if ($param[0] eq "hspnumfilter"){
            $inputsql=$param[1];
     }
```

```
if ($param[0] eq "colour"){
            $colour=$param[1];
    }
  }

  print "sqlscore is $inputsql; blastfile is $tid; artout is $artout; var is $variable; colour is
$colour\n";
  open (OUT, ">$artout") or die "Cannot open the output file: $artout \n";
  my $DB_Type="production";
  my $Oracle_ID="version2";
  my $Oracle_Passwd="version2";
  my $dbh = getConnection($DB_Type, $Oracle_ID, $Oracle_Passwd);
  my $flen=3+(length $variable)+13;
  my $strpad=0;
  my $emptystring="";
  while($strpad<$flen){
     $emptystring=$emptystring." ";
     $strpad++;
  }

  my $sth = $dbh->prepare("select blast_run_id from v2_blast_run where
blast_output_file_name=\'$tid\'");
  $sth->execute();
  my @maxarr=$sth->fetchrow_array();
  my $bid=shift @maxarr;
  if($bid){

    $sth = $dbh->prepare("select distinct query_id, query_update_date, target_id,
target_update_date from v2_blast_hsp where blast_run_id=$bid");
    $sth->execute();
    my @allhits;
    while(@allhits=$sth->fetchrow_array()){
            my $q=shift @allhits;
            my $qd=shift @allhits;
            my $t=shift @allhits;
            my $td=shift @allhits;
       my $sth4 = $dbh->prepare("select db_accession_num, sequence_description,
sequence_length from v2_database_sequence where database_sequence_id=$t and
seq_update_date=\'$td\'");
            $sth4->execute();
       my @seq2=$sth4->fetchrow_array();
            $sth4->finish();

            my $sth2 = $dbh->prepare("select sequence_name from v2_database_sequence where
database_sequence_id=$q and seq_update_date=\'$qd\'");
```

```
          $sth2->execute();
          my @seq=$sth2->fetchrow_array();
          print "dbaccnum is $seq2[0]; seqname is $seq[0]; description is $seq2[1]; length is
$seq2[2]\n";
          my @bias=split(/\|/, $seq[0]);
          my $start;
          my $end;
          my $strand;
          if($seq[0]=~/ORF/){
            $strand=$bias[7];
            $start=$bias[5];
            $end=$bias[6];
          }
          else{
            $strand=$bias[6];
            $start=$bias[4];
            $end=$bias[5];
          }
          $sth2->finish();
          print "biasstart is $start; biasend is $end\n";
          #my $strand="";
          my $expect="";
          my $score="";
      my $geneline="";
          my $percents="";
          my $sth1 = $dbh->prepare("select hsp_query_start, hsp_query_end, hsp_strand,
percent_identity, score, expect from v2_blast_hsp where blast_run_id=\'$bid\' and query_id=$q
and query_update_date=\'$qd\' and target_id=$t and target_update_date=\'$td\' $inputsql");
          $sth1->execute();
          my @hitpos;
          my $count=0;
          while(@hitpos=$sth1->fetchrow_array()){
            my $newstart=($hitpos[0]*3)-1+$start;
            my $newend=($hitpos[1]*3)-1+$start;
            print "hsp_start is $hitpos[0]; hsp_end is $hitpos[1]; newstart
is $newstart; newend is $newend\n";
            #$strand=$hitpos[2];
            $score=$hitpos[4];
            $expect=$hitpos[5];
            if($percents){
                $percents=$percents.",$hitpos[3]";
        }
        else{
                $percents="$hitpos[3]";
        }
```

```
if($geneline){
                $geneline=$geneline.",$newstart..$newend";
        }
        else{
                $geneline="$newstart..$newend";
        }
            $count++;
     }
         $sth1->finish();
         if($geneline){
        print "$geneline at genenum $count and strand $strand\n";
        if($count>1){
           $geneline="join(".$geneline.")";
        }
        if($strand eq "plus"){
             print OUT "FT   $variable            $geneline\n";
        }
        elsif($strand eq "minus"){
           print OUT "FT   $variable            complement($geneline)\n";
        }
            else{
          .....die "invalid strand of $strand\n";
            }
            print OUT "FT".$emptystring."/blast_score=$expect\n";
            print OUT "FT".$emptystring."/colour=$colour\n";
            print OUT "FT".$emptystring."/percent_id=$percents\n";
            print OUT "FT".$emptystring."/score=$score\n";

        print OUT "FT".$emptystring."/subject_id=$seq2[0]\n";
        print OUT "FT".$emptystring."/note=\"$seq[0]\"\n";
            print OUT "FT".$emptystring."/note=\"$seq2[1]\"\n";
            print OUT "FT".$emptystring."/note=\"$seq2[2] total bp\"\n";
     }
   }
 }
 else{
    die ("invalid blast output file name\n");
 }
  $sth->finish();
  $dbh->disconnect();
}
```