Acetyl-CoA in plant biology

Heather Leigh Babka
Iowa State University

Follow this and additional works at: https://lib.dr.iastate.edu/etd

Part of the Cell and Developmental Biology Commons, and the Genetics and Genomics Commons

Recommended Citation
https://lib.dr.iastate.edu/etd/10870

This Dissertation is brought to you for free and open access by the Iowa State University Capstones, Theses and Dissertations at Iowa State University Digital Repository. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Iowa State University Digital Repository. For more information, please contact digirep@iastate.edu.
Acetyl-CoA in plant biology

by

Heather Leigh Babka

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Major: Genetics

Program of Study Committee:
Eve Syrkin Wurtele
Diane Bassham
Harry T. Horner
Basil J. Nikolau
David Oliver

Iowa State University
Ames, Iowa
2009

Copyright © Heather Leigh Babka, 2009. All rights reserved.
# TABLE OF CONTENTS

## CHAPTER 1: GENERAL INTRODUCTION
- Dissertation Organization 1
- Acetyl-CoA Metabolism 1
- Bioactive compounds of *Hypericum gentianoides* are found in schizogenously formed translucent glands 4
- References 7
- Figures 11

## CHAPTER 2: PHENOTYPE REVERSION IN ATP-CITRATE LYASE MUTANTS SUPPLEMENTED WITH MALONATE IS NOT REGULATED BY BIOSYNTHETIC TRANSCRIPTS
- Abstract 12
- Introduction 13
- Results 16
- Discussion 29
- Materials and Methods 38
- References 41
- Figures 48

## CHAPTER 3: *HYPERICUM GENTIANOIDES* PRODUCES NOVEL BIOACTIVE COMPOUNDS IN SCHIZOGENOUSLY FORMED GLANDS
- Summary 73
- Introduction 74
- Results 78
- Discussion 88
- Summary 97
- Experimental Procedures 99
- References 106
- Figures 113

## CHAPTER 4: GENERAL CONCLUSIONS

## APPENDIX A: EXPLORASE: MULTIVARIATE EXPLORATORY ANALYSIS AND VISUALIZATION FOR SYSTEMS BIOLOGY

## APPENDIX B: METNETDB: A PLANT BIOLOGICAL NETWORK DATABASE BASED ON A LABELED GRAPH MODEL
APPENDIX C. ATP-CITRATE LYASE TARGETED TO PLASTIDS OF *ARABIDOPSIS THALIANA* IS UNDER COMPLEX POST-TRANSCRIPTIONAL CONTROL

ACKNOWLEDGEMENTS
CHAPTER 1: GENERAL INTRODUCTION

Dissertation Organization

This dissertation contains a study about acetyl-CoA metabolism in Arabidopsis thaliana and describes the storage and production of specialized acetyl-CoA-derived polyketide phytochemicals in Hypericum gentianoides. Chapter 2 focuses on the transcriptomic and metabolomic impacts of reducing the activity of ATP-citrate lyase in Arabidopsis and the subsequent rescue of these plants by the application of malonate. Chapter 3 centers on structural studies of H. gentianoides translucent glands, which contain bioactive polyketides. Chapter 4 closes the dissertation. This dissertation also contains three appendices. Appendix A is a manuscript about a novel visualization and statistical analysis computer program exploRase that has been published in the Journal of Statistical Software, and that I have used to analyze transcriptomic and metabolomic data. A manuscript that is being submitted to Plant Physiology about the MetNetDB database and its labeled graph structure is featured in Appendix B. Appendix C discusses several challenges in over-expressing ACL in the plastids of Arabidopsis.

Acetyl-CoA Metabolism

Acetyl-CoA is critical for the formation of many important secondary biomolecules in plants, including isoprenoids, anthocyanins, and elongated fatty acids. The CoA moiety is quite large and it is thought to be unable to pass though biological membranes, thus acetyl-CoA must be synthesized within the compartment in which it is going to be utilized (Brooks and Stumpf, 1966). ATP-citrate lyase produces the cytoplasmic pool of acetyl-CoA via the cleavage of citrate to form oxaloacetate and acetyl-CoA in an ATP-dependent manner.
The cytosolic pool of acetyl-CoA can be utilized in the three different ways: carboxylation, condensation, or acetylation (Figure 1). Carboxylation reactions involving cytosolic acetyl-CoA include the elongation of fatty acids (Pollard and Stumpf, 1980), the synthesis of cuticular waxes (Xu et al., 2002), and the production of flavonoids and anthocyanins (Hrazdina et al., 1978). The condensation of acetyl-CoA in the cytosol to form acetoacetyl-CoA leads to the production of mevalonate-derived isoprenoids (sesquiterpenes, sterols, and brassinosteroids), the initial steps occurring in a combination of cytosolic (Disch et al., 1998) and peroxisomal (Reumann et al., 2007; Sapir-Mir et al., 2008) pathways. Acetylation of several types of molecules, including a subset of isoprenoids (Shalit et al., 2003), sugars (Pauly and Scheller, 2000), anthocyanins (Bloor and Abrahams, 2002), phenolics (Shalit et al., 2003; Whitaker and Stommel, 2003), and alkaloids, occurs in a variety of subcellular locations. The primary source of plastidic acetyl-CoA is the pyruvate dehydrogenase complex (pPDC) (Ke et al., 2000). Acetyl-CoA can also be produced via plastidic acetyl-CoA synthetase (ACS); this enzyme is thought to prevent acetate toxicity and/or to shunt carbon into the glyoxylate cycle (Lin and Oliver, 2008). The mitochondrial pyruvate dehydrogenase complex (mPDC) is the primary source of mitochondrial acetyl-CoA (Randall et al., 1981; Givan and Hodgson, 1983). Acetyl-CoA is produced in microbodies through β-oxidation breakdown of lipids. Histones within the nucleus are acetylated, but it is not known how the acetyl-CoA for this process is produced in the nucleus.
ATP-Citrate Lyase

ACL is the enzyme responsible for generating the pool of cytosolic acetyl-CoA in Arabidopsis (Fatland et al., 2002). ACL activity has also been detected in a number of other plant species including: Pisum sativum (Kaethner and ap Rees, 1985), Glycine max (Nelson and Rinne, 1975) and Brassica napus (Ratledge et al., 1997). Two subunits compose the ACL heteromer in Arabidopsis: ACLA and ACLB (Fatland, 2002). It is likely the Arabidopsis ACL is an octoheteromer having an A₄B₄ configuration (Fatland, 2002). The Arabidopsis genome contains three genes that encode ACLA (ACLA-1, At1g10670; ACLA-2, At1g60810; ACLA-3, At1g09430) and two genes encoding ACLB (ACLB-1, At3g06650; ACLB-2, At5g49460) (Fatland et al., 2002; Fatland et al., 2005). The expression of ACL subunit transcripts are coordinated in both space and time in flower development, early seed development and developing trichomes (Fatland, 2002). The expression of ACL subunits is also coordinated with the expression of homomeric acetyl-CoA carboxylase, the enzyme responsible for utilizing acetyl-CoA to elongate fatty acids in the cytosol. Experiments in yeast have shown that ACL activity is present only when both ACLA-1 and ACLB-2 polypeptides are present. Yeast expressing only one subunit of ACL (ACLA or ACLB) lack ACL activity (Fatland, 2002).

Arabidopsis plant lines having an antisense copy of ACLA-1 inserted into their genome (antisense-ACLA plants) and with reduced ACL activity, hyper-acumulate starch and anthocyanins, have altered cuticular wax, and have a small, dark green, bonsai-like appearance (Fatland et al., 2005). The severity of the phenotype is correlated to the degree of reduction in ACL activity. Given the complex phenotype observed in antisense-ACLA
plants, it is clear that cytosolic acetyl-CoA generated by ACL is critical to normal
development of Arabidopsis. Antisense-ACLA plants fed malonate showed a reversion of the
high starch and anthocyanin phenotype, as well as normal growth and development. This
indicates that the phenotype observed in antisense-ACLA plants is related to deficiencies in
the carboxylation branch of cytosolic acetyl-CoA metabolism.

Antisense-ACLA and wildtype (WT) were treated with malonic acid or water and
subjected to transcriptomic and metabolic analyses in order to further understand the global
impacts of reduced cytosolic acetyl-CoA pool.

**Bioactive compounds of Hypericum gentianoides are found in schizogenously formed
translucent glands**

*The genus Hypericum*

*Hypericum* has approximately 450 members and belongs to the Clusiaceae family.
Several species of *Hypericum* are known for producing compounds with pharmacological
activity with *H. perforatum* (also known as St. John’s Wort) being the most widely studied
species. Several classes of metabolites that have been found in *Hypericum* species include
monoterpenes, diterpenes, triterpene sterols, phenylpropanoids, aglycone flavonoids,
glycosylated flavonoids, biflavones, benzophenones, xanthones, acyl-phloroglucinols,
anthrones, dianthrones and porphorins (Ernst, 2003).

*Hypericum perforatum* has been shown to have anti-inflammatory (Hammer et al.,
2007), antibacterial (Franklin et al., 2009), antiviral (Axarlis et al., 1998), and antidepressive
activity (Mennini and Gobbi, 2004). Two compounds isolated from *H. perforatum* have been
heavily studied: hypericin and hyperforin. Hypericin accumulates in the black glandular nodules of *H. perforatum* (Ciccarelli et al., 2001) and has been linked to antiviral activity (Meruelo et al., 1988). Hypericin is associated with photosensitivity in unpigmented areas of skin in animals that have ingested it (Ernst, 2003). Hypericin and pseudohypericin extracted from *H. perforatum* have significant light dependent cytotoxicity in cultured cells (Schmitt et al., 2006). Hyperforin, an acyl-phloroglucinol, is thought to be responsible for the anti-depressive activity of *H. perforatum* (Mennini and Gobbi, 2004). Hyperforin accumulates in translucent glands found on *H. perforatum* (Soelberg et al., 2007). While hypericin and hyperforin are associated with *Hypericum* species, not all members of the species accumulate these two compounds (Crockett et al., 2005). One of these species, *H. gentianoides*, accumulates several unusual acyl-phloroglucinols, saroaspidin A, uliginosin A and hyperbrasilol C that have been shown to have bioactive properties (Hillwig, 2008).

*Hypericum secretory structures*

Gland development can proceed in a schizogenous and/or lysigenous manner. In schizogenously developing glands, the central lumen is formed as secretory cells pull apart. During lysigenous gland development secretory cells go through programmed cell death and secretory material is released as the cells dissolve. Studies have suggested oil cavities found on the fruits of *Citrus medica*, *C. limon*, and *Poncirus trifoliata* develop lysigenously, while oil glands on *Citrus sinensis* (Fahn, 1979) and gum resin ducts of *Rhus glabra* (Fahn and Evert, 1974) appear to develop schizogenously. Secretion of material in secretory tissues can be classified as holocrine (the substance is released from the secretory cells as result of
cellular disintegration) or merocrine (the substance is eliminated/excreted from the intact cytoplasm) (Fahn, 1979). Merocrine secretion occurs one of two ways: (1) ecrine secretion which occurs when a concentration gradient or active process exists for the elimination of secretory substances; or (2) secretory substance can be collected in vesicles and expelled from the cytoplasm as in granulocrine secretion (Fahn, 1979).

Species of *Hypericum* often have prominent secretory glands (translucent and/or dark) on leaves, sepals, and petals (Ernst, 2003). Translucent glands contain cells that secrete an essential oil (Ernst, 2003), which in *H. perforatum* includes hyperforin (Soelberg et al., 2007). Dark glands contain a waxy substance filled with hypericin that is responsible for staining skin red when the gland bursts open (Ernst, 2003). The developmental stages of translucent and dark glands are different. Dark glandular nodules do not look like any other type of internal secretory structure observed in other plants (Curtis and Lersten, 1990). Secretory cells contained within dark nodules remain intact and grow in size as they accumulate secretory product, but unlike laticifers, the material is never secreted at the leaf surface (Onelli et al., 2002). Dark glandular nodules also lack a central lumen that is typically found in secretory cavities and ducts (Onelli et al., 2002). Similar observations have been made for dark glandular nodules found on *H. richeri* (Maffi et al., 2003). Larger peripheral cells surrounding black nodules have been linked to increased hypericin and pseudohypericin production, indicating they are site of biosynthesis (Kornfeld et al., 2007).

*Hypericum perforatum* translucent glands and secretory canals can be further classified into types based on where they appear on the plant and how they form (Ciccarelli et al., 2001). Type A secretory canals are found in floral tissues as well as in stem and root cortical parenchyma, and they are delimited by four polygonal cells. Type B secretory
canals, found on leaves and sepals, are similar to the translucent glands. Secretory canals that develop using a combination of schizogeny and lysigeny are found on the ovary and style are classified as Type C. The translucent glands of *H. perforatum* develop schizogenously (Curtis and Lersten, 1990; Ciccarelli et al., 2001). Translucent glands from *H. richeri* also have been studied, and while they appear to develop like *H. perforatum* glands, they never develop into mature secretory cavities (Maffi et al., 2003). On the other hand, translucent glands found on *H. balearicum* do not appear to develop schizogenously or lysigenously, but via a different method altogether (Curtis and Lersten, 1990).

*Hypericum gentianoides* plants only have translucent glands and we wanted to investigate whether they developed in a manner similar to observations made for *H. perforatum* or more like the pattern of development observed for *H. balearicum*.

Understanding the development of these glands is important since they are likely the site of synthesis and/or accumulation of saroaspidin A, uliginosin A and hyperbrasilol C.

**References**


Fatland BL (2002) ATP-citrate lyase has an essential role in cytosolic acetyl-CoA production in Arabidopsis Iowa State University, Ames, IA
Franklin G, Conceição LFR, Kombrink E, Dias ACP (2009) Xanthone biosynthesis in Hypericum perforatum cells provides antioxidant and antimicrobial protection upon biotic stress. Phytochemistry 70: 60-68
Hillwig ML (2008) Understanding the biological activity of Hypericum species through metabolomics studies. Iowa State University


Figure 1. Acetyl-CoA metabolism network. The reaction of ACL is shown in the center. The acetyl-CoA synthesized via ACL can then be carboxylated to produce elongated fatty acids, flavonoids, and cuticle as shown in purple. Cytosolic acetyl-CoA can also be used for acetylation of various metabolites as shown in green. Isoprenoids and sterols are produced via the condensation of cytosolic acetyl-CoA as shown in blue. Mevalonate-derived isoprenoids are synthesized in the cytosol and peroxisome (Sapir-Mir et al., 2008). The enzyme responsible for emodinanthrone synthesis is a type III polyketide synthase, but it has not been isolated. CHS – Chalcone Synthase, BPS – Benzophenone Synthase, BUS – Isobutyrophene Synthase
CHAPTER 2. PHENOTYPE REVERSION IN ATP-CITRATE LYASE MUTANTS SUPPLEMENTED WITH MALONATE IS NOT REGULATED BY BIOSYNTHETIC TRANSCRIPTS

A paper to be submitted to Plant Cell

Heather L. Babka1, Suh-Yeon Choi1, Di Cook2, Beth L. Fatland1, Hilal Ilarslan1, Basil J. Nikolau3, Eve Syrkin Wurtele1

ABSTRACT

Cytosolic acetyl-CoA provides the necessary building blocks for the biosynthesis of a number of important cytosolic-derived biomolecules including, flavonoids, sterols, and elongated fatty acids. Arabidopsis plants with reduced levels of the enzyme that supplies acetyl-CoA to the cytosol, ATP-citrate lyase (ACL), have a severe phenotype characterized by the hyper-accumulation of starch and anthocyanins, and severely reduced cuticular wax. Malonic acid treatment of such antisense-ACLA mutants alleviates the high starch and anthocyanin content, and renews cuticular wax accumulation to a wildtype (WT) level. Transcriptomic and metabolomic experiments were carried out using antisense-ACLA and WT plants treated with either water or malonate. Genes of starch synthesis and flavonoid synthesis exhibit increased expression in antisense-ACLA mutants. However, expression of these genes is unaffected by the application of malonic acid, even though levels of starch and flavonoids accumulation and the overall phenotype of these plants revert to WT. Only 0.21% of transcripts are significantly affected by treatment of antisense-ACLA plants with malonate. Similarly, only 0.97% of the detected metabolites show altered levels of accumulation after malonate rescue of antisense-ACLA plants. The reversal of the complex phenotype exhibited

1 Department of Genetics, Development, and Cell Biology; Iowa State University; Ames, IA
2 Department of Statistics; Iowa State University; Ames, IA
3 Department of Biochemistry, Biophysics, and Molecular Biology; Iowa State University; Ames, IA
by antisense-ACLA plants appears to be the result of a number of post-transcriptional changes and surprisingly few transcriptional changes.

**INTRODUCTION**

Acetyl-CoA is an intermediate in a number of metabolic processes essential to the normal operation of plant cells. Acetyl-CoA cannot cross membranes, and must be synthesized within the cellular compartment in which it is to be utilized (Brooks and Stumpf, 1966; Fatland et al., 2005; Oliver et al., 2009). The plastidic pool of acetyl-CoA is primarily synthesized by the plastidal pyruvate dehydrogenase (Lin and Oliver, 2008) and is utilized primarily in *de novo* fatty acid biosynthesis. Acetyl-CoA in the mitochondria is synthesized via the mitochondrial form of pyruvate dehydrogenase complex, which is then fed into the citric acid cycle (Mooney et al., 2002). Peroxisomal/glyoxysomal acetyl-CoA is derived from β-oxidation of fatty acids stored either as storage compounds or as products of membrane degradation under stress, and feeds into the glyoxylate shunt (Oliver et al., 2009) and a subset of isoprenoids (Sapir-Mir et al., 2008). ATP-citrate lyase (ACL) synthesizes the plant’s cytosolic pool of acetyl-CoA in an ATP-dependent manner from citrate, forming oxaloacetate and acetyl-CoA (Kaethner and ap Rees, 1985; Fatland et al., 2002). Cytosolic acetyl-CoA is utilized in a number of biochemical reactions via carboxylation, condensation, or acetylation reactions (Figure 1). The carboxylation pathway of cytosolic acetyl-CoA metabolism gives rise to elongated fatty acids (Pollard and Stumpf, 1980) utilized in the formation of some seed oils, some membrane phospholipids, the sphingolipid ceramide moiety (Sperling and Heinz, 2003), cuticular waxes (Fatland et al., 2005), suberin (Kolattukudy, 1980), flavonoids and anthocyanins (Hrazdina et al., 1978; Fatland et al.,
2005) and malonyl-derivatives (e.g. D-amino acids and malonylated flavonoids) (Fatland et al., 2005). The condensation of acetyl-CoA leads to the formation of mevalonate-derived isoprenoids such as sterols and brassinosteroids (Newman and Chappell, 1999); the condensation process occurs in the cytosol and, as shown by recent localization (Sapir-Mir et al., 2008) and proteomics (Reumann et al., 2007) studies, in the peroxisomes. In addition a subset of proteins (Houtz et al., 2008), isoprenoids (Shalit et al., 2003), sugars (Pauly and Scheller, 2000), anthocyanins (Bloor and Abrahams, 2002), phenolics (Shalit et al., 2003; Whitaker and Stommel, 2003) and alkaloids can be acetylated by acetyl-CoA.

ACL has been characterized in plant species including Arabidopsis (Fatland et al., 2005), Pisum sativum (Kaethner and ap Rees, 1985), Glycine max (Nelson and Rinne, 1975), and Brassica napus (Ratledge et al., 1997). The Arabidopsis ACL protein is a heteromer encoded by two subunits: ACLA is similar to the N-terminal end of mammalian ACL, while ACLB is similar to the C-terminus (Fatland et al., 2002). ACL is likely an octoheteromer of the configuration A_4B_4 (Fatland et al., 2002).

Arabidopsis plants expressing an antisense-ACLA transgene have reduced ACL enzyme activity, along with aberrations in growth, development and composition (Fatland et al., 2005). ACL activity of about 50% of the wildtype (WT) level leads to plants with a severe phenotype, characterized by a dark green bonsai-like appearance, hyper-accumulation of anthocyanins and starch, and severe reductions in specific products of cytosolic acetyl-CoA, such as cuticular waxes and seed coat constituents. A preliminary evaluation of transcript accumulation using a cDNA partial-genome array (Newman et al., 1994) showed shifts in the accumulation of transcripts of starch and anthocyanin biosynthesis was associated with the antisense-ACLA genotype (Fatland et al., 2005).
In order to identify what pathways might be associated with the antisense phenotype; antisense-ACLA plants were supplied with a variety of compounds known to require cytosolic acetyl-CoA for their biosynthesis. Of the substrates tested, malonate was unique at alleviating the antisense-ACLA phenotype and causing reversion to WT levels of anthocyanin, starch and cuticular wax (Fatland et al., 2005).

The complex phenotypic changes induced in plant lines expressing the antisense-ACLA gene would be likely associated with concurrent global shifts in transcript and metabolite accumulation. Preliminary evidence shows the transcriptome is indeed distinct (Fatland et al., 2005). Because the antisense-ACLA lines revert to a WT-like visual phenotype following treatment with malonate, it also might be expected that transcript accumulation in these plants would revert to a near WT signature. Similarly, the metabolome might become more WT-like in antisense-ACLA plants treated with malonic acid, since malonyl-CoA would be less limiting, thus allowing the plant to carry out normal metabolic functions.

To gain a better understanding of the perturbations caused by reduced cytosolic acetyl-CoA levels, we use a coupled transcriptomic and metabolomics approach to compare antisense-ACLA and WT plants. Ten days after malonate or water treatment, plants were photographed, weighed, and collected for microarray and metabolomics analysis. Here, we identify large changes in starch, anthocyanin, and fatty acid biosynthetic genes in antisense-ACLA compared to WT plants treated with water. The transcriptome and metabolome of antisense-ACLA plants was evaluated after reversion of plants to a visually normal phenotype following treatment with malonate. Our surprising conclusion is that a reversion to a near
WT appearance by antisense-ACLA plants treated with malonate results in few major transcriptional or metabolomic changes.

RESULTS

Expression of the ACL genes.

Accumulation of ACLA and ACLB subunit transcripts is both spatially and temporally coordinated during development of flowers, seeds, and trichomes; in these, the ACL subunit genes are co-expressed in the same tissues and cells as the homomeric ACCase (Fatland et al., 2002). This may reflect that ACL is providing the cytosolic acetyl-CoA substrate used by homomeric ACCase for fatty acid elongation and formation of other malonyl-derivatives (Fatland et al., 2002).

In order to understand the co-expression of ACL subunits over a variety of conditions, we evaluated the co-accumulation of each of the five ACL subunit transcripts across a range of genetic and environmental perturbations as well as development, using MetaOmGraph software (MetNetDB.org). In a data set from 70 microarray experiments (about 1000 chips) from public databases (Mentzen et al., 2008; Mentzen and Wurtele, 2008) transcripts from two of the three ACLA genes and the two ACLB genes are co-accumulated (Figure 2 A). For example, ACLA-1 and ACLB-2 transcripts co-accumulate across a variety of conditions, in particular in fruit and the male gametophyte (Figure 2 A).

To get a better idea of whether ACL transcripts were co-expressed with genes involved in fatty acid biosynthetic processes we looked at the Regulon classification of ACL transcripts in Arabidopsis ((Mentzen and Wurtele, 2008). (A Regulon can be defined as a group of genes that are co-expressed over a number of environmental and genetic conditions...
(Kim et al., 2001; Stuart et al., 2003); approximately half of Arabidopsis genes can be considered as members of a Regulon (criteria include a 0.7 or greater correlation across 1000 microarray samples, (Mentzen and Wurtele, 2008). ACLB-2 is a member of co-expression Regulon 167, which contains a total of nine genes. Four members of Regulon 167 are implicated in fatty acid-related processes: an acyl carrier protein (ACP1, At3g05020), an acyl-ACP thioesterase (FatA, At3g25110), ACLB-2 (At5g49460), and biotin carboxyl carrier protein isoform 2 (BCCP2, At5g15530). Other genes in Regulon 167 are: a pyrophosphate-dependent phosphofructokinase (At1g76550), a hydroxymuconic semialdehyde hydrolase (At4g24140), an NC-domain containing protein (At1g01225), and two genes with unknown functions (At2g16760, At4g12700). The expression of Regulon 167 members is highest in developing seeds (Ohlrogge and Browse, 1995) and male gametophyte (Evans et al., 1992), two organs that tissues produce large amounts of fatty acids. These data provide a further indication of a possible relationship of ACL expression and fatty acid elongation. The other ACL subunit transcripts are not members of the Regulons described by Mentzen and Wurtele (2008).

Phenotype of WT and antisense-ACLA mutants collected for transcriptomic and metabolomic studies

To evaluate in more detail the metabolites and transcripts that are altered in antisense-ACLA mutants, as well as to determine how the accumulation of these molecules is affected when malonic acid is supplied and the morphological and developmental phenotype reverts to WT-like (Fatland et al., 2005), we used a combined non-targeted metabolomics and transcriptomic approach. Plants were grown in Magenta boxes in a random block design.
under continuous illumination, treated with 0.01 mM malonic acid or water at 16 days after imbibition (DAI), and grown for an 10 additional days (Fatland et al., 2005). As described in Fatland et al. (2005) antisense-ACLA plants have a small, bonsai-like growth habit (Figure 3 B) and hyper-accumulate anthocyanins and starch (Figure 3 F) as compared to WT plants (Figure 3 A, E). When malonate is applied to antisense-ACLA plants, plants become WT-like with respect to growth habit (Figure 3 D), development, starch accumulation (Figure 3 H), and anthocyanin accumulation (Fatland et al., 2005). Malonic acid does not affect the morphology, development or levels of starch and anthocyanin accumulation of WT plants (Figure 3 C, G).

For each biological replicate, whole rosettes (without roots or inflorescences) were collected, pooled (3 – 4 plants/replicate), and frozen in liquid nitrogen. Rosettes were pulverized using a liquid nitrogen-cooled mortar and pestle and the resultant powder was allocated for transcriptomic and metabolomic analyses. At least four biological replicates were analyzed for each treatment.

**Analysis of transcriptomic data**

*Statistical analysis of transcriptomic data by traditional analysis.* Transcriptomic data is often analyzed using a statistical test to get p-values, followed by p-value correction to control the false discovery rate, resulting in q-values. Lists of genes having altered gene expression are created using a q-value cutoff (i.e., “accept all genes as differentially expressed if q-value is below 0.01”). Genes with similar expression patterns can be grouped by clustering.
Analysis of transcriptomics data shows large differences in transcriptome of untreated WT and antisense-ACLA mutants

Antisense-ACLA mutants that had been exposed to water-only (untreated with malonic acid) were compared to WT plants given water-only. The antisense-ACLA mutants had major alterations in 172 transcripts, using criteria of p-value < 0.04 and a q-value cutoff of 0.30. In total 55 transcripts accumulate to lower levels in antisense-ACLA–H₂O compared to WT-H₂O, whereas 117 transcripts accumulate to higher levels in antisense-ACLA-H₂O plants (Supplemental Table I).

Limma/exploRase analysis of transcriptomics data reveals additional expression differences

We were able to take advantage of our experimental design to implement a limma/exploRase based data analysis approach (Swayne et al., 2003; Cook et al., 2007; Lawrence et al., 2008) to evaluate changes in transcript profiles. Interestingly, no genes were found to be differentially expressed using a T-test or ANOVA in comparisons of antisense-ACLA significantly changed in antisense-ACLA –H₂O compared to antisense-ACLA-MA). Therefore, we designed and implemented a limma/exploRase-based analysis that relies on combined statistical analysis and graphical visualization. The general approach is as follows. A statistical test is used to calculate p-values (similar to many microarray analyses). Transcripts identified by this test to be differentially accumulated are then examined visually and placed in user-defined categories based on accumulation patterns; thus a reliance on q-values is eliminated (Swayne et al., 2003; Cook et al., 2007; Lawrence et al., 2008). Once transcripts having interesting accumulation patterns are identified, additional transcripts are
identified using Pearson’s correlation. ExploRase has the ability to search for user defined patterns within the data using the “Find Specific Pattern” function. This method may provide a useful general method to analyze transcriptomic data not amenable to usual statistical analysis.

To implement this approach, microarray data from the four genotype-by-treatment combinations were subjected to limma analysis (Smyth, 2005) within the program exploRase (Lawrence et al., 2008) to calculate F-statistics and p-values. Those transcripts having an F-statistic > 10 (p-value < 0.0065) are included in the limma/exploRase analysis and on the gene list in supplemental table II (http://vrac.iastate.edu/~hbabka/).

Given the design of the microarray experiment, twelve expression patterns are possible for any transcript that is differentially expressed (Figure 4). The graphs provide an alternative way to visualize data, so that it is easier to observe accumulation differences among the four genotype-by-treatment combinations. Each potential pattern is designated according to the transcript accumulation change from one genotype-by-treatment combination to another.

In order to identify transcripts that exhibit a particular accumulation profile, the “Find Pattern” function of exploRase was used (Wurtele et al., 2003; Lawrence et al., 2008). Each gene falling within the cutoff (p-value < 0.0065) was visualized using exploRase graphics to ensure expression was significantly altered in the sample medians, as well as ensuring that the replicates for each genotype-by-treatment condition were consistent. Differentially-accumulating transcripts were identified for five of the twelve possible patterns (Figure 4).
These are:

- up in antisense-ACLA,
- down in antisense-ACLA,
- up in antisense-ACLA-MA,
- up in antisense-ACLA-H₂O
- down in antisense-ACLA-H₂O.

Unusual aspects of the genes with differentially-accumulating transcripts are characterized in the ensuing results sections. (Supplemental Table SII provides the entire list of the genes with differentially-accumulating transcripts; http://vrac.iastate.edu/~hbabka/).

Genotype comparisons (WT all treatments versus antisense-ACLA all treatments) show a total of 505 transcripts are differentially accumulating at an F-statistic > 10 (p-value < 0.0065). In contrast, despite the reversion to WT appearance of antisense-ACLA plants treated with malonate, far fewer transcripts were significantly changed in antisense-ACLA-H₂O-treated plants compared to antisense-ACLA-MA (48 transcripts at an F-statistic > 10, p-value < 0.0065).

Transcript accumulation associated with genotype (WT all treatments versus antisense-ACLA all treatments).

Transcripts encoding starch and anthocyanin biosynthesis and, fatty acid modification/elongation are up-accumulated in the antisense-ACLA genotype regardless of treatment. In total, 300 transcripts have a higher accumulation in antisense-ACLA mutants, regardless of treatment, as compared to WT plants. Of these, 147 have functional assignments in TAIR.
Notably, four of these transcripts appear to be involved in starch biosynthesis: two large subunits of ADP-glucose pyrophosphorylase (APL4, AT2G21590; APL3, AT4G39210); a putative starch synthase (AT1G32900); and a putative sucrose synthase (AT3G43190) (Figure 5A). When malonic acid is applied to antisense-ACLA plants, the starch over-accumulation phenotype disappears (Figure 3H). However, the expression of these four starch biosynthesis genes is high in antisense-ACLA plants, regardless of treatment or level of starch accumulation.

Antisense-ACLA plants accumulate large amounts of anthocyanin pigments, however this phenotype is alleviated by the application of malonic acid (Fatland et al., 2005). Seven anthocyanin biosynthesis genes are up-regulated in the antisense-ACLA plants, regardless of malonic acid treatment (Figure 5B): chalcone synthase (CHS, AT5G13930), which catalyzes the first committed step of flavonoid biosynthesis via the condensation of three molecules of malonyl-CoA with a molecule of p-coumaryl-CoA; leucoanthocyanidin dioxygenase (LDOX, AT4G22880), which converts dihydroflavonols into flavonols (Stracke et al., 2009); anthocyanidin synthase (ANS, At4G22870), responsible for the conversion of leucoanthocyanidin to colored anthocyanidin (Davies and Schwinn, 2006); flavanone 3-hydroxylase (F3H, AT3G51240), the third enzyme in central flavonoid metabolism, which has been shown to physically interact with CHS (Owens et al., 2008); flavonol 3-O-glucosyltransferase-like (AT5G54060); dihydroflavonol 4-reductase (DFR, AT5G42800); the glutathione S-transferase TRANSPARENT TESTA 19 (TT19, AT5G17220), mutants of which lack the ability to take up anthocyanidin 3-glucosides into the vacuole (Kitamura et al., 2004). These flavonoid biosynthesis transcripts, like the starch biosynthesis transcripts,
accumulate to high levels in antisense-ACLA plants, but their accumulation changes little upon application of malonic acid.

Five transcripts encoding long chain fatty acid metabolism enzymes accumulate to higher levels in antisense-ACLA plants as compared to WT plants (Figure 5C). These are: *DAISY* (AT1G04220), a β-ketoacyl-CoA synthase that plays a key role in the synthesis of aliphatic monomers >20 C for use in suberin (Franke et al., 2009); *LACS1* (AT2G47240), a long-chain-fatty-acid-CoA ligase who substrates are palmitate (16:0), palmitoleate (16:1Δ9), and linolate (18:2), and to a lesser extent, stearate (18:0) and oleate (18:1) (Shockey et al., 2002), with the highest preference for C30:0 very long chain fatty acids (Lu et al., 2009)- the LASCs family esterify free fatty acids to acyl-CoA’s, activating them for further metabolism; (*FAD3*, AT2G29980), one of three omega-3 fatty acid desaturases that convert linoleic acid to α-linoleic acid- a jasmonic acid precursor and a potential detoxifier of stress-induced reactive oxygen species (Mene-Saffrane et al., 2009); a putative fatty acid elongase (*KCS-16*, AT4G34250); and an aldehyde decarbonylase (*CER1*, AT1G02205). Even though malonate treatment of antisense-ACLA plants has been shown to restore the cuticular wax deposition to WT-like levels (Fatland et al., 2005), the expression of these five fatty acid related genes is not altered upon application of malonate.

Three other transcripts that accumulate preferentially in antisense-ACLA plants are: cyclin delta-3 (*CYCD3*, AT4G34160), a cellulose synthase (*CEL1*, AT1G70710), and glutamine dumper 1 (*GDU1*, AT4G31730), each involved in tasks associated with cell size. (Figure 5D). Antisense-ACLA plants have smaller cells as compared to WT and this difference is alleviated by malonate (Fatland et al., 2005). *CYCD3* expression increases as the cell cycle starts and *CYCD3* mutants have reduced numbers of cells in the leaf, but
unaltered leaf size (Dewitte et al., 2007). Cell growth and cell wall loosening are thought to occur via CEL1, which is expressed in actively growing tissues, especially thickening walls (Shani et al., 2006). When GDU1 is over-expressed, increased amounts of glutamine are excreted through the leaf hydathodes and the plants are smaller compared to (Pilot et al., 2004).

We used a Fisher’s exact test to identify whether any Arabidopsis Regulons (Mentzen and Wurtele, 2008) are significantly over-represented in genes having the highest expression in antisense-ACLA plants. Thirteen Regulons are over-represented in the up in antisense-ACLA genes (Table 1) Eighty-one of the differentially expressed genes on the up in antisense-ACLA list are members of Regulon 4, the over-represented Regulon (p-value=2 X 10^{-16}) in which mitosis–associated genes predominate. Regulon 36, which contains genes primarily for lipid modification and cuticular wax synthesis, is also highly over-represented (p value = 1 X 10^{-6}).

A Fisher’s exact test was carried out using the “common pathway” tool at MetNetDB (www.metnetdb.org). Several MetNet pathways are over-represented (p-value < 0.05) in genes having the highest expression in antisense-ACLA and they include pathways involved in anthocyanin, flavonoid, starch, and epicuticular wax biosynthesis (Table 2).

*Transcripts that are under-accumulated in antisense-ACLA (antisense-ACLA all treatments versus WT all treatments).*

205 transcripts accumulate to lower levels in antisense-ACLA plants as compared to WT regardless of treatment. Five transcripts that are down in antisense-ACLA-H2O are shown in Figure 6: a putative pyruvate decarboxylase (PDC4, AT5G01320), ACLA-3
(AT1G09430), a putative cysteine synthase (AtcysD1, AT3G22460), ACLA-2 (AT1G60810), and anthranilate synthase beta subunit (TRP4, AT1G25220). These particular transcripts were of interested because of their involvement in primary metabolism, including the biosynthesis of acetyl-CoA (in multiple subcellular compartments) and amino acid metabolism. Pyruvate decarboxylase is one of the components of the pyruvate dehydrogenase complex that operates in the chloroplast and mitochondria (Lin et al., 2003), and it is hypothesized that crosstalk occurs between subcellular locations (Lunn, 2007). Arabidopsis has four pyruvate decarboxylase genes; of these, PDC4 is expressed in several tissues (Kursteiner et al., 2003). AtcysD1 is one of the cytosolic cysteine synthases in Arabidopsis; it is thought the expression of the AtcysD1 isoform is dependent on the environment and/or plant part (Yamaguchi et al., 2000). TRP4 catalyzes the first committed step of tryptophan biosynthesis (Niyogi et al., 1993).

Within this down-in-antisense-ACLA gene list, a total of 10 Regulons are over-represented (Table 3). Three of these Regulons (10, 24 and 46) are associated with defense responses.

MetNet pathways over-represented with the down-in-antisense-ACLA gene list include; the acetyl-CoA biotin network and several amino acid biosynthetic pathways (Table 4).

**Transcript accumulation associated with malonic acid treatment.**

*Transcripts having a genotype-by-treatment interaction effect.* 48 transcripts accumulate differentially in antisense-ACLA water-treated plants versus antisense-ACLA malonate-
treated plants. This observation was surprising in light of the visual phenotypic reversion of ACL plants following treatment with malonic acid.

*Transcripts having up in antisense-ACLA-MA pattern of expression.* In total, five transcripts accumulate to higher levels in antisense-ACLA plants that had been treated with malonate, relative to the other three samples (Figure 7). Only two of these transcripts have been annotated: response regulator 16 (*ARR16*, AT2G40670), which has been shown to respond to cytokinin signaling in the roots of Arabidopsis (Kiba et al., 2002); and a membrane-anchored ubiquitin-fold protein (*AtMUB5*, AT1G77870), a class of proteins that are involved in post-translational modification of proteins (Downes et al., 2006).

*Transcripts with higher levels of accumulation in water-treated antisense-ACL plants.* Twenty-six transcripts accumulate to the highest level in antisense-ACLA-H₂O plants. All of the transcripts on this list returned to WT levels after treatment with malonic acid.

Three of the four transcripts from the thalianol biosynthesis operon over-accumulate in water-treated antisense-ACLA mutants, relative to the WTs and the malonate-treated antisense-ACLA plants. The thalianol operon (Field and Osbourn, 2008) is responsible for producing the triterpenoid thalianol, and contains four genes: AT5G47980 (BADH acyltransferase), AT5G47990 (thaliana-diol desaturase, THAD), AT5G48000 (thalianol hydroxylase, THAH), and AT5G48010 (thalianol synthase, THAS). The accumulation of *THAD*, *THAH*, and *THAS* transcripts are 2-fold higher in antisense-ACLA plants treated with water and the level of these transcripts returns to WT levels upon malonic acid treatment of the antisense-ACLA mutants (Figure 8). As shown in the GoTree layout in Figure 9, the GO terms for this operon are over-represented (p-value = 0.0215).
Three other examples involved in cell growth and development are shown in Figure 10: *SEPALLATA 2* (*SEP*, AT3G02310), an expansin (*ATEXPA10*, AT1G26770), and *APPETALA 3* (*AP3*, AT3G54340). *ATEXPA10*, encodes an expansin that has control over cell size by causing cell wall loosening (Cho and Cosgrove, 2000), whereas *SEP2* and *AP3* are floral identity genes (Pelaz et al., 2000; Malcomber and Kellogg, 2005).

Within the highest-in-antisense-ACLA-H2O gene list, MetNet pathways for glycolipid biosynthesis (p-value = 0.0039) and jasmonic acid (p-value = 0.017) signaling are over-represented.

*Transcripts with lower levels of accumulation in the antisense-ACLA mutant phenotype.*

Figure 11 shows two transcripts that accumulate to lowest levels in antisense-ACLA-H2O mutants, a beta-ketoacyl-CoA synthase family protein (AT1G07720), and a sulfate transporter known as AST91 (AT1G23090). In total, 18 transcripts have this pattern of expression.

**Metabolomics analysis**

Metabolites were extracted from a minimum of 4 replicates for each genotype/treatment, and separated into polar and nonpolar fractions (Roessner et al., 2001). After replicates were analyzed by GC-MS, the resultant peaks were integrated and aligned using Metabolomics Ion-based Data Extraction Algorithm (MET-IDEA) (Broeckling et al., 2006). The peak areas were log2 transformed and mean centered, prior to limma (Smyth, 2005) analysis within exploRase (Lawrence et al., 2008). In total 31 nonpolar and 72 polar metabolite peaks were observed (Supplemental Table SIII; http://vrac.iastate.edu/~hbabka/).
limma/exploRase analysis of metabolite accumulation in antisense-ACLA and WT plants

Using the limma/exploRase approach described for transcriptomics data, metabolites were categorized accumulation pattern (Figure 4). Four polar metabolites accumulate at higher levels in antisense-ACLA plants regardless of treatment; a chemical identity has not been assigned to any of these (Figure 12 A). A nonpolar metabolite, also not chemically identified, accumulates to higher levels in antisense-ACLA-H2O (Figure 12 B). Serine accumulation is lower in antisense-ACLA plants treated with water, but returns to WT-levels in antisense-ACLA plants treated with malonate. (Figure 12 C).

Additional differentially accumulated metabolites were identified by comparisons within a single genotype (antisense-ACLA vs. WT) or treatment (MA or H2O) (Table 1). An unknown polar metabolite and the nonpolar metabolite, docosanoic acid, accumulate to higher levels, and two hexoses and glucose accumulate to lower levels, in H2O-treated antisense-ACLA plants compared to malonic acid-treated antisense-ACLA plants. In comparisons of malonic acid-treated WT to malonic acid-treated antisense-ACLA, trans-sinapinic acid accumulates to lower levels in antisense-ACLA, while an unknown polar metabolite is higher in antisense-ACLA. In comparisons of H2O-treated WT and antisense-ACLA plants, an oligosaccharide accumulates to higher levels in H2O-treated antisense-ACLA plants. There are no significant changes in the accumulation of metabolites in WT plants treated with H2O compared to WT treated with malonic acid.
DISCUSSION

Most transcriptional changes in antisense-ACLA are not reverted by malonate.

Transcriptome data indicates that the dramatic reversions observed in the antisense-ACLA plants phenotype after treatment with malonic acid are likely not dependent on concomitant changes in transcript accumulation of the biosynthetic genes in the affected pathways. Indeed, treatment with malonic acid alters the accumulation of only 48 transcripts in antisense-ACLA plants. Transcripts in starch and flavonoid biosynthesis are up-regulated in the antisense-ACLA plants, however, this increase is not reversed with malonate treatment despite reversion of the starch and flavonoid hyper-accumulation phenotype.

Starch biosynthetic transcript accumulation is higher in antisense-ACLA plants regardless of treatment

Antisense-ACLA plants hyper-accumulate the transcripts of several starch synthesis genes, including the regulatory subunits of ADP-glucose pyrophosphorylase (Crevillen et al., 2005), APL3 and APL4. While antisense-ACLA plants treated with malonic acid contain WT-like levels of starch, accumulation of transcripts involved in starch synthesis is unchanged by malonate treatment. Thus, although malonic acid abolishes the hyper-accumulation of starch, the signal inducing APL3 and APL4 transcript over-accumulation is still present. It is thought that accumulation of sucrose leads to an increase in the level of trehalose-6-phosphate in the cytosol, which in turn activates ADP-glucose pyrophosphorylase and shunts photosynthate from sucrose into starch (Lunn, 2007). Transcription of both APL3 and APL4, regulatory subunits of ADP-glucose
pyrophosphorylase, has been shown to be induced by sucrose and trehalose (Crevillen et al., 2005). This may mean sucrose and/or trehalose concentrations are higher within antisense-ACLA plants as compared to WT. Thus, transcription of APL3 and APL4 is induced to higher levels within antisense-ACLA plants regardless of treatment. The changes observed in antisense-ACLA-MA (reversion of starch hyper-accumulation phenotype) may reflect possible post-translational modification to ADP-glucose pyrophosphorylase in response to a starch synthesis inhibitor such as Pi (Zeeman et al., 2007), whereas increased sucrose and/or trehalose continue to induce gene expression.

Flavonoid biosynthetic transcripts accumulate to high levels in antisense-ACLA plants regardless of treatment

A large battery of flavonoid biosynthesis genes show enhanced expression in antisense-ACLA plants, including CHS, the first committed step of flavonoid synthesis, and F3H the third step in the pathway (Owens et al., 2008); DFR, TT19, and ANS, LDOX and a flavonol-3-O glucosyl transferase are also up-regulated in anti-sense ACLA plants. Concentrations of acetyl-CoA in WT Arabidopsis leaves are about 0.10 nmol/g fresh weight (Perera et al., 2009). Surprisingly, antisense-ACLA plants divert what may be very low levels of cytosolic acetyl-CoA into the production of anthocyanins (Fatland et al., 2005). These pigments are synthesized in response to a number of environmental stresses including UV light exposure (Chalker-Scott, 1999) and abnormal temperatures (Havaux and Kloppstech, 2001), and their increase might indicate that antisense-ACLA plants sense that they are under stress. Flavonoids also have been shown to regulate auxin transport (Mathesius et al., 1998;
Buer and Muday, 2004; Buer et al., 2007), and it is possible that the altered anthocyanin content in the antisense-ACLA plants participates in some aspect of signaling.

It is surprising that antisense-ACLA plants would shunt the cytosolic acetyl-CoA they do have available into the flavonoid pathway (Figure 13). One possible explanation for the usage of cytosolic acetyl-CoA in flavonoid biosynthesis, is that antisense-ACLA plants are making large amounts of anthocyanins as a defense mechanism against a perceived attack, perhaps to an environmental signal; this is consistent with the overrepresentation of genes from regulon 9 (response to environmental stimuli) in antisense-ACLA plants. Transcripts from \textit{CHS}, \textit{F3H}, \textit{DFR}, \textit{ANS} and \textit{TT19} accumulate under nutrient stresses including; phosphorous and nitrogen deficiency, high light intensity and high sucrose (Lillo et al., 2008). Certainly, one could imagine that insufficient cytosolic acetyl-CoA levels would create stress that would be similar to other forms of abiotic stress. Presumably, the glutathione S-transferase \textit{TT19} (AT5G17220) is up-regulated in antisense-ACLA in order to deal with the import of the excess anthocyanins into the vacuole. Mutants of \textit{TT19} lack the ability to import anthocyanidin 3-glucosides into the vacuole (Kitamura et al., 2004). While antisense-ACLA plants that have been treated with malonic acid contain WT-like amounts of anthocyanin, the expression of these genes does not decrease upon malonic acid treatment. This indicates that the reduction of anthocyanin pigments in malonic acid-treated antisense-ACLA plants is not related to changes in transcript accumulation of these seven anthocyanin synthesis genes, but is facilitated by another mechanism. Perhaps, jasmonate levels are higher in antisense-ACLA as compared to WT plants regardless of treatment; jasmonate induces expression of flavonoids biosynthetic genes. \textit{CHS}, \textit{F3H}, \textit{DFR}, and \textit{LDOX} expression are all induced in the presence of jasmonate plus sucrose (Loreti et al., 2008).
Alternatively, three Regulons representing defense responses (10, 24, and 46) are down regulated in antisense-ACLA plants, even though the increased amounts of anthocyanin indicate antisense-ACLA plants are responding to some type of stress. Regulon 10 contains transcripts involving defense response that is mediated through protein modification. Regulon 46 includes has 29 genes associated with the synthesis of protective compounds derived from shikimate (Mentzen and Wurtele, 2008). Shikimate is utilized in the synthesis of several aromatic amino acids (tyrosine, phenylalanine, and tryptophan) as well as ligins, flavonoids, and indole containing molecules. These results indicate that carbon flux is being diverted from the shikimate pathway, into other pathways including starch and flavonoid biosynthesis.

**Wax biosynthetic transcripts are up-regulated in antisense-ACLA plants regardless of treatment.**

Fatland et al. (2005) observed that antisense-ACLA plants have vastly decreased cuticular wax; this decrease in wax content could be associated with the over-accumulation in antisense-ACLA plants of the battery of transcripts involved in altering and elongating fatty acids. *FAD3*, a ER localized fatty acid desaturase, is involved in the production of α-linolenic acid (18:3) from linoleic acid (18:2) on phosphatidylcholine (Browse et al., 1993), a precursor of the defense compound jasmonic acid, and an antioxidant that reacts with and detoxify reactive oxygen species (Mene-Saffrane et al., 2009). Arabidopsis plants deficient in FAD3 activity do not have the ability to produce jasmonic acid (McConn and Browse, 1996). *DAISY*, a long-chain acyl-CoA synthase, is involved in the production of very long chain fatty acid derivatives (C > 20) found in the aliphatic domain of suberin (Franke et al.,
2009), another indication that some defense responses may be up-regulated in antisense-ACLA plants. It is interesting to note, in light of the decreased cuticular wax in antisense-ACLA plants, that the CER1 wax synthesis gene is up-regulated in antisense-ACLA plants. This may be a response to the reduced levels of C29 alkanes and C29 secondary alcohols in antisense-ACLA plants (Fatland et al., 2005). Mutations in CER1 result in Arabidopsis with a reduced ability to convert C30 aldehydes to C29 alkanes (Aarts et al., 1995) and CER1 is thought to produce odd-numbered carbon length alkanes and secondary alcohols (Samuels et al., 2008). LACS1 encodes a novel acyl-CoA synthetase that is active in the synthesis of cuticular wax and of cutin monomers (Lu et al., 2009). LACS1 may over-expressed in antisense-ACLA plants as a response to the altered levels of cuticular wax. The overall decrease in cuticular wax accumulation in antisense-ACLA (Fatland et al., 2005), is reflected in the altered transcription of genes in regulon 36 (lipid modification and cuticular wax synthesis). As noted, malonic acid treatment of antisense-ACLA reverts the cuticular wax to WT levels, but does not alter the over-accumulation transcripts of these cuticular wax synthesis genes (Fatland et al., 2005).

Docosanoic acid accumulates to higher levels in antisense-ACLA plants receiving water than those receiving malonic acid. The product of the DAISY gene (AT1G04220) has been shown to encode an active docosanoic acid synthase (Franke et al., 2009) and is one of the transcripts that accumulates in the up in antisense-ACLA pattern. Interestingly, the level of DAISY expression does not decrease in antisense-ACLA plants treated with malonic acid, but the product of the resultant enzyme does. It is also interesting that docosanoic acid accumulates to higher levels in antisense-ACLA plants since it is a long-chain fatty acid, elongated using cytosolic malonyl-CoA. One would expect that malonate-treated plants
would have more malonyl-CoA available to be used for fatty acid elongation. Alternately, *DAISY* has been shown to respond to salt, drought stress, and wounding (Franke et al., 2009), so the increased expression of *DAISY* and its end product docosanoic acid could be a response to stress that the lack of cytosolic acetyl-CoA causes in the plant.

**Transcripts accumulating to lower levels in antisense-ACLA plants regardless of treatment**

As expected, the expression of ACLA subunits is decreased in antisense-ACLA plants in response to the insertion of the antisense copy of *ACL-A1*. The decreased expression of *PDC4* and *TRP4* may be in response to the lack of carbon available for other pathways in antisense-ACLA plants.

**Sulfur metabolism is decreased in antisense-ACLA plants regardless of treatment**

Serine and cysteine are involved in the assimilation of sulfur in Arabidopsis (Kopriva, 2006). The decreased *AtcysD1* and *AST19* transcripts, as well as the lowered accumulation of the amino acid serine, indicate that sulfur metabolism may be affected in antisense-ACLA plants. Interestingly, both *AST19* mRNA and serine accumulate to lower levels in antisense-ACLA-H$_2$O, and return to WT levels following treatment with malonic acid. In photosynthetic tissues serine is synthesized from products of the glyoxylate cycle or from 3-phosphoglycerate (Ho and Saito, 2001) in the plastid. This may indicate that changes in acetyl-CoA levels in the cytosol affects amino acid metabolism in plastids.
Transcripts from the thalianol operon accumulate to higher levels in antisense-ACLA-H₂O plants

Thalianol synthesis operon transcripts are enriched in antisense-ACLA-H₂O, highest in antisense-ACLA-H₂O, and lowest in antisense-ACLA-MA. The thalianol synthesis operon is up-regulated in antisense-ACLA plants and the increase is reversed by the application of malonic acid. This triterpenoid has only been reported in produced in roots (Field and Osbourn, 2008), and it will be interesting to determine whether thalianol is also evident in rosettes of antisense-ACLA-H₂O plants.

Development and cell cycle transcripts accumulate to higher levels in antisense-ACLA-H₂O plants

The expansin, ATEXPA10, controls cell size by causing cell wall loosening (Cho and Cosgrove, 2000). ACL plants have smaller cells than WT plants (Fatland et al., 2005). The small cell size phenotype is alleviated by the application of malonic acid, thus ATEXPA10 is an example of a transcript whose accumulation pattern matches the phenotypic observations.

It is difficult to explain the altered expression of SEP2 and AP3. SEP2 and AP3 are floral identity genes, with SEP2 expressed in all parts of the flower, while AP3 is necessary for the formation of petals and stamens (Pelaz et al., 2000; Malcomber and Kellogg, 2005). It has been reported that the Gerbera SEP2 homolog (GhGRCD2) is expressed in leaf tissue (Malcomber and Kellogg, 2005). Perhaps SEP2 and AP3 are expressed in leaves under special circumstances, such as the one created when Arabidopsis has decreased ACL activity. In fact antisense-ACLA plants have a longer life span and flowers that have shorter than
normal petals (Fatland et al., 2005), which could be caused by the altered expression of SEP2 and AP3 in leaf tissues.

ACL plants have a reduced ability to synthesize cytosolic acetyl-CoA, which is essential for the production of many key metabolites (Figure 1). While malonate reverses the severe phenotype of ACL plants (starch content, anthocyanin content, and cuticular wax morphology), the reversal is not reflected in the transcriptomic data.

**Sterol biosynthesis may be reduced in antisense-ACLA plants**

A number of sterol biosynthesis mutants have a small plant size phenotype. Co-suppression of AdoMet-Sterol-C24-methyl transferase (SMT2) (Schaeffer et al., 2001) induces a morphology similar to antisense-ACLA plants, including increased branched, a protruding pistil during flowering, prolonged growth, and low fertility. SMT2 is an enzyme that influences the level of campesterol by controlling conversion of cycloartenol to 24-alkyl-Δ5-sterols (Figure 13), which leads to the production of stigmasterol (Schaeffer et al., 2001). In each mutant, the phenotype is rescued by treatment with sitosterol, but not with brassinolide (Schaeffer et al., 2001; Fatland, 2002). In each mutant, a key enzyme in sitosterol synthesis is eliminated causing alteration in the ratio of campesterol: sitosterol. Taken together, these results indicate that sitosterol deficiency might be responsible for some aspects of the antisense-ACLA phenotype. The severe phenotype observed in antisense-ACLA plants could be caused by an interplay among decreased sitosterol levels, large amounts of anthocyanin which do not inhibit auxin transport (Peer and Murphy, 2007) and the sequestration of available carbon in starch (Figure 13). It is possible that together, alteration of the levels of these three biomolecules (starch, anthocyanin, and sitosterol)
contribute to some aspects of the extreme phenotypic changes that can not be explained by global changes to gene expression and metabolite accumulation alone.

**Phenotype reversion of antisense-ACLA plants by malonate can not be explained by transcriptome changes alone**

In combination, these data indicate that much of the phenotype reversion observed in antisense-ACLA-MA plants can not be explained by transcriptome changes alone. One interpretation of our data is that many of the large-scale changes in antisense-ACLA-MA plants are being carried out at the protein level, for example, through regulation of translation or enzyme activity. Alternatively, antisense-ACLA-MA plants may be reallocating limited resources in an efficient manner that creates changes that are too small to accurately detect. Similar phenomena have been observed in water-starved maize, in which ovary abortion has been induced (Boyer and McLaughlin, 2007). When water-starved ovaries are compared to normal ovaries, large changes are observed at the transcriptome level using cDNA microarrays (Zinselmeier et al., 2002), but when sucrose is fed to maize stems causing a partial phenotype reversion, significantly fewer genes show altered transcript accumulation (McLaughlin and Boyer, 2004). Certainly, depriving a plant of a key metabolite or resource causes complex changes in the genome, transcriptome, proteome, and metabolome. When the alteration is caused at the genome level, the simple application of a chemical will not repair the causal perturbation, so while the phenotype can be reverted, the original mutation still exists. Thus it becomes difficult to dissect the causal phenotypic effects from secondary phenotypes. In the case of antisense-ACLA plants, it is difficult to determine which observed phenotype is indicative of acetyl-CoA deficiency and which phenotype(s) are secondary.
Within antisense-ACLA plants it is likely that an intricate interplay involving sugar metabolism, anthocyanin biosynthesis, auxin signaling, and altered fatty acid metabolism merge to produce a complex phenotype.

**MATERIALS AND METHODS**

**Plant Materials and Growth**

*Arabidopsis thaliana* ecotype Columbia transformed with antisense-ACLA (Fatland et al., 2005) and the WT (WT) Columbia ecotype were used in these studies. Plants were grown in Magenta boxes containing 60 mL of MS media, in a growth room under continuous illumination (170 μmol m⁻² s⁻¹) at 22°C as previously described (Fatland et al., 2005).

At 16 DAI, each seedling was assessed for phenotype, and randomly treated with 0.25 mL of either filter-sterilized water or 0.6 mM malonate as previously detailed (Fatland et al., 2005). Ten days treatment (at 26 DAI) plants were photographed, weighed, pooled (3 – 5 rosettes per pool), and frozen and store in liquid nitrogen for microarray and metabolomics analysis. Less than 2 min occurred between harvest and submersion in liquid nitrogen. The same samples were used in transcriptomic and metabolomic analysis.

At 26 DAI, 10 plants from each treatment were stained for starch with IKI (Berlyn et al., 1976). Seedlings were incubated in 95 % ethanol for 48 h to remove pigments and then incubated overnight in a solution containing 1% potassium iodide (KI) and 1% iodine (I). Seedlings were rinsed with water for 30 min and photographed.
Microarray Analysis

RNA was extracted using trizol (AFGC). The resultant RNA was further purified using a Qiagen RNeasy Micro kit, diluted to 1µg/µL and submitted to the Iowa State University GeneChip Facility (Ames, IA) for labeling, hybridization, and scanning of the Affymetrix ATH1 arrays.

In total there were 4 replicates of WT treated with water, 5 replicates of WT treated with malonic acid, 4 replicates of antisense ACL treated with water, and 4 replicates of antisense ACL treated with malonic acid.

The raw data were normalized using R (http://www.r-project.org/) via the bioconductor (http://www.bioconductor.org/) package RMA (Irizarry et al., 2003). The normalized data were further analyzed in the R-based GUI, exploRase (metnetdb.org; Lawrence et al., 2008) using the integrated limma analysis (Smyth, 2005) interface and pattern searching tool. When a pattern of interest was observed, additional transcripts with this pattern of accumulation were found via correlation using the “Find Similar Entities” function. Transcripts with an F-statistic < 10 tended to have small changes in gene expression, so this value was selected as a cutoff.

Gene lists were submitted to AtGeneSearch (metnetdb.org) to get current annotation data. For each gene list, overrepresented Gene Ontology (GO) terms were obtained using GOstat (Beissbarth and Speed, 2004), organized using GOTree (Nick Ransom, personal communication), and graphed using Cytoscape (Shannon et al., 2003).

Metabolic profiling and quantification

Samples were extracted using the methods from Bligh and Dyer (1959) and Roessner
(2001) with modifications. Samples were homogenized in liquid nitrogen with 20 μL of internal standards (polar: adipic acid and ribitol, 1mg/mL in H2O; nonpolar: nonadecanoic acid, 2 mg/mL in chloroform), transferred to a glass tube containing 2 mL of methanol, and placed at 70°C for 15 min. Then 0.5 mL of H2O and 1 mL of chloroform were added and samples were vortexed for 1 min. Additional 0.5 mL of H2O was added to separate polar and nonpolar phase. The lower nonpolar phase was methylated by adding 2 mL of 1N hydrochloric acid in methanol and incubated 90°C for 1 hr. After extracting with 4 mL of H2O twice and dehydratated by addition of anhydrous sodium sulphate, nonpolar extracts were filtered through PTFE (pore size 2 μm) filter (Supelco, Bellefonte, PA) and completely dried using nitrogen gas. The upper polar phase was filtered through a PTFE (pore size 2 μm, high protein binding) filter (Supelco, Bellefonte, PA) and completely dried using a speedvac system (Savant instruments, Farmingdale, NY). The dried polar extracts were methoxymated by 20 mg/mL of methoxyamine-HCl in pyridine at 30°C for 90 min, and both polar and nonpolar extracts were silylated by N,O-bis(trimethylsilyl) trifluoroacetamide with 1% trimethylchlorosilane at 37°C for 1 hr.

The GC/MS analysis was performed in W. M. Keck Metabolomic research Lab at Iowa State University (Ames, IA). A GC series 6890 from Agilent (Palo Alto, CA) coupled with 5973 Agilent mass detector, using a Capillary HP-5 (30m X 0.25 mm id, film thickness 0.25 um) column and helium carrier gas was used. The column temperature was programmed to increase from 70°C to 260°C at a rate of 5°C / min, then held for 10 min, and again up to 320°C at a same rate. Detected Mass (m/z) range was 50-750, and total running time was 60 min.
DATA processing and integration

Chromatograms were integrated by AMDIS (Automated mass spectral deconvolution and identification system). Peaks were integrated and aligned using MET-IDEA (Metabolomics Ion-based Data Extraction Algorithm) (Broeckling et al., 2006), (htexperimentsle.org/PlantBio/MS/MET-IDEA/index.html). The resultant data were log2 transformed and mean-centered in R (http://www.r-project.org). The mean-centered metabolomics data was further analyzed in the R based GUI exploRase (metnetdb.org) to find significantly differentially accumulating metabolites using limma analysis (Smyth, 2005) and visual pattern assessment (Lawrence et al., 2008).

ACKNOWLEDGMENTS

We would like to extend our special thanks to Dr. Ann Perera at the W. M. Keck Metabolomics Laboratory, at Iowa State University for providing the analytical instrumentation; and Drs. Jiqing Peng and Steve Whitham at the GeneChip® Facility at Iowa State University for processing of the microarrays. This work was supported by the following grants: National Science Foundation Arabidopsis 2010s: DBI 0520267 and MCB-0820823.

REFERENCES


Lin, M., Behal, R., and Oliver, D.J. (2003). Disruption of plE2, the gene for the E2 subunit of the plastid pyruvate dehydrogenase complex, in Arabidopsis causes an early embryo lethal phenotype. Plant Molecular Biology 52, 865-872.


Figure 1. Acetyl-CoA metabolism network. The ACL reaction occurs in the cytosol, creating a cytosolic pool of acetyl-CoA. Cytosolic acetyl-CoA can be carboxylated to produce elongated fatty acids, flavonoids, and cuticle (maroon). Cytosolic acetyl-CoA can be the substrate for acetylation of various metabolites in (green). Predominantly mevalonate-derived isoprenoids such as sterols and sesquiterpenes are produced via the condensation of acetyl-CoAs (blue), these condensation reactions occur in both the peroxisome and cytosol (Reumann et al., 2007; Sapir-Mir et al., 2008), thus either pool of acetyl-CoA might contribute to these reactions.
Figure 2

![Graph showing gene expression levels](image)

<table>
<thead>
<tr>
<th>Gene</th>
<th>ACLA-1</th>
<th>ACLA-2</th>
<th>ACLA-3</th>
<th>ACLB-1</th>
<th>ACLB-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACLA-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACLA-2</td>
<td>44.49%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACLA-3</td>
<td>-13.59%</td>
<td>-19.70%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACLB-1</td>
<td>47.98%</td>
<td>46.70%</td>
<td>-23.02%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACLB-2</td>
<td>62.79%</td>
<td>62.71%</td>
<td>-8.97%</td>
<td>57.54%</td>
<td></td>
</tr>
</tbody>
</table>
Figure 2 continued

<table>
<thead>
<tr>
<th>Gene</th>
<th>At1g01225</th>
<th>At1g76550</th>
<th>At2g16760</th>
<th>ACP1</th>
<th>FatA</th>
<th>At4g12700</th>
<th>At4g24140</th>
<th>BCCP2</th>
<th>ACLB-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>At1g01225</td>
<td>53.86%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At1g76550</td>
<td>33.87%</td>
<td>66.42%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At2g16760</td>
<td>62.10%</td>
<td>72.24%</td>
<td>71.30%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACP1</td>
<td>74.86%</td>
<td>73.62%</td>
<td>53.59%</td>
<td>79.20%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FatA</td>
<td>59.00%</td>
<td>56.08%</td>
<td>44.08%</td>
<td>73.90%</td>
<td>70.38%</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At4g12700</td>
<td>21.16%</td>
<td>43.02%</td>
<td>87.48%</td>
<td>59.99%</td>
<td>40.34%</td>
<td>38.00%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>At4g24140</td>
<td>47.82%</td>
<td>63.24%</td>
<td>75.60%</td>
<td>87.11%</td>
<td>64.58%</td>
<td>75.16%</td>
<td>70.72%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BCCP2</td>
<td>33.04%</td>
<td>60.33%</td>
<td>57.48%</td>
<td>70.45%</td>
<td>49.62%</td>
<td>47.86%</td>
<td>42.35%</td>
<td>62.10%</td>
<td></td>
</tr>
<tr>
<td>ACLB-2</td>
<td>62.10%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 2. Accumulation of ACL transcripts. Transcript data from 70 microarray experiments from public databases (Mentzen and Wurtele, 2008) viewed using MetaOmGraph (MetNetDB.org). Panel A. Co-expression of the three ACLA (ACLA-1, At1g10670; ACLA-2, At1g60810; ACLA-3, At1g09430) transcripts and two ACLB (ACLB-1, At3g06650; ACLB-2, At5g49460) transcripts. The expression of ACLA-1 and ACLB-2 in yeast is sufficient for ACL activity in yeast (Fatland et al., 2002). The table contains Pearson’s correlations calculated for each transcript. Panel B. Expression of transcripts found in regulon 167 (Mentzen and Wurtele, 2008). Four members of regulon 167 (Mentzen and Wurtele, 2008) are involved in fatty acid related processes and they are: an acyl carrier protein (ACP1, At3g05020), an acyl-ACP thioesterase (FatA, At3g25110), ACLB-2 (At5g49460), and biotin carboxyl carrier protein isoform 2 (BCCP2, At5g15530). Other genes in regulon 167 include: an NC-domain containing protein (At1g01225), pyrophosphate-dependent phosphofructokinase (At1g76550), a hydroxymuconic semialdehyde hydrolase (At4g24140), and two unknowns (At2g16760 and At4g12700). As shown above the expression of these genes spikes in fruit and male gametophytes, tissues having high fatty acid demands. The table contains Pearson’s correlations calculated for each transcript. In correlation tables: dark orange, correlation 80 – 90 %; pink, correlation 70 – 80 %; yellow, correlation 60 – 70 %; and gray, correlation 100 %.
Figure 3. Phenotypes of WT and antisense-ACLA plants. Panels A – D: WT and antisense-ACLA plants treated with water or malonic acid, just prior to harvest. Panels E – H: WT and antisense-ACLA plants treated with water or malonic acid were stained for starch using IKI (Berlyn et al., 1976).
Significantly different transcripts and metabolites were categorized based on the 12 possible accumulation patterns. Given the design of the microarray experiment, twelve expression patterns are possible for each transcript. A line connects WT-MA and ACL-H₂O to make it easier to observe accumulation differences. A total of 553 differentially-accumulated transcripts were identified; these fell into only five patterns (up in antisense-ACLA, down in antisense-ACLA, up in antisense-ACLA MA, up in antisense-ACLA-H₂O, down in antisense-ACLA-H₂O), red line (p value < 0.0065). Six metabolites differentially accumulate in one of the possible patterns (up in antisense-ACLA, up in antisense-ACLA-H₂O, down in antisense-ACLA-H₂O) using a p-value < 0.05 as a cut-off.
Figure 5.
Figure 5 continued.
**Figure 5. Seventeen transcripts up in antisense-ACLA plants.** A total of 300 transcripts are up in antisense-ACLA plants. Parallel coordinate plots showing median gene expression data (logged, RMA normalized) for WT-H2O treated, WT-malonic acid (MA) treated, antisense-ACLA-H2O treated, and antisense-ACLA-MA treated plants. **Panel A:** Expression of starch synthesis transcripts, large subunit of ADP-glucose pyrophosphorylase 4 (APL4, AT2G21590) (Crevillen et al., 2005), large subunit of ADP-glucose pyrophosphorylase 3 (APL3, AT4G39210) (Crevillen et al., 2005), a putative starch synthase (AT1G32900), and a putative sucrose synthase (AT3G43190). **Panel B:** Expression of flavonoids biosynthesis genes, chalcone synthase (CHS, AT5G13930), a leucoanthocyanidin dioxygenase (LDOX, AT4G22880) (Stracke et al., 2009), flavanone 3-hydroxylase (F3H, AT3G51240) (Owens et al., 2008), flavonol 3-O-glucosyltransferase-like (AT5G54060), and dihydroflavonol 4-reductase (DFR, AT5G42800) (Yuan et al., 2007). **Panel C:** Expression of fatty acid genes, beta-ketoacyl-CoA synthase (DAISY, AT1G04220) (Franke et al., 2009), a long-chain-fatty-acid-CoA ligase (LACS1, AT2G47240) (Shockey et al., 2002), an omega-3 fatty acid desaturase (FAD3, AT2G29980) (Mene-Saffrane et al., 2009), a fatty acid elongase (AT4G34250), and an octadecanal decarbonylase (CER1, AT1G02205) (Aarts et al., 1995). **Panel D:** Expression of cyclin delta-3 (CYCD3, AT4G34160) (Dewitte et al., 2007), a cellulase (CELI, AT1G70710) (Shani et al., 2006), and glutamine dumper 1 (GDU1, AT4G31730) (Pilot et al., 2004).
Table 1. Regulons over-represented in up in antisense-ACLA accumulating transcripts.

300 transcripts are up in antisense-ACLA plants. The table shows the number of significantly expressed genes in the up in antisense-ACLA pattern also occurring in listed Regulons. The total number of transcripts in each Regulon is also shown.

1 (Mentzen and Wurtele, 2008)

2 Fisher’s exact test was used to calculate p-values.

<table>
<thead>
<tr>
<th>Regulon(^1) - postulated physiological function</th>
<th># over-expressed genes / Total # genes in</th>
<th>p-value(^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 - mitosis</td>
<td>81/571</td>
<td>2.20 \times 10^{-16}</td>
</tr>
<tr>
<td>36 - lipid modification and cuticular wax synthesis (flowers and shoot apex - specific)</td>
<td>6/43</td>
<td>1.26 \times 10^{-6}</td>
</tr>
<tr>
<td>77</td>
<td>4/18</td>
<td>6.54 \times 10^{-6}</td>
</tr>
<tr>
<td>89</td>
<td>6/15</td>
<td>1.71 \times 10^{-5}</td>
</tr>
<tr>
<td>146</td>
<td>3/10</td>
<td>2.00 \times 10^{-4}</td>
</tr>
<tr>
<td>95</td>
<td>5/15</td>
<td>2.00 \times 10^{-4}</td>
</tr>
<tr>
<td>297</td>
<td>2/5</td>
<td>0.002</td>
</tr>
<tr>
<td>321</td>
<td>3/5</td>
<td>0.002</td>
</tr>
<tr>
<td>252</td>
<td>3/6</td>
<td>0.004</td>
</tr>
<tr>
<td>235</td>
<td>3/7</td>
<td>0.007</td>
</tr>
<tr>
<td>109</td>
<td>3/13</td>
<td>0.009</td>
</tr>
<tr>
<td>380</td>
<td>2/3</td>
<td>0.012</td>
</tr>
<tr>
<td>359</td>
<td>1/5</td>
<td>0.036</td>
</tr>
</tbody>
</table>

\(^1\) (Mentzen and Wurtele, 2008)
Table 2. MetNet pathways over-represented in up in antisense-ACLA accumulating transcripts. The up in antisense-ACLA gene list was entered into the “common pathway” tool to find over-represented pathways. The table shows the most significantly over-represented MetNet pathways occurring in the up in antisense-ACLA transcripts.

1 www.metnetdb.org

2 Fisher’s exact test was used to calculate p-values

<table>
<thead>
<tr>
<th>MetNet Pathway</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>anthocyanin biosynthesis</td>
<td>0.001</td>
</tr>
<tr>
<td>leucopelargonidin and leucocyanidin biosynthesis</td>
<td>0.002</td>
</tr>
<tr>
<td>superpathway of fatty acid biosynthesis</td>
<td>0.005</td>
</tr>
<tr>
<td>starch biosynthesis</td>
<td>0.008</td>
</tr>
<tr>
<td>epicuticular wax biosynthesis</td>
<td>0.009</td>
</tr>
<tr>
<td>hydroxycinnamic acid tyramine amides biosynthesis</td>
<td>0.018</td>
</tr>
<tr>
<td>flavonoid biosynthesis</td>
<td>0.026</td>
</tr>
<tr>
<td>choline biosynthesis I</td>
<td>0.027</td>
</tr>
<tr>
<td>sucrose degradation (v4.0)</td>
<td>0.029</td>
</tr>
<tr>
<td>AGRIS regulatory network - full</td>
<td>0.031</td>
</tr>
<tr>
<td>superpathway of glyoxylate cycle</td>
<td>0.033</td>
</tr>
</tbody>
</table>
Figure 6. Five transcripts down in antisense-ACLA plants. 205 transcripts are down in antisense-ACLA plants. Parallel coordinate plots show median gene expression data (logged, RMA normalized) for WT-H_2O treated, WT-MA treated, antisense-ACLA-H_2O treated, and antisense-ACLA-MA treated plants. The redundancy of ACLA-2 and ACLA-3 transcripts is expected, because the antisense construct was designed to down-regulate both of these. Putative pyruvate decarboxylase (PDC4, AT5G01320) (Lin et al., 2003); ACLA-3 (AT1G09430) (Fatland et al., 2002), a putative cysteine synthase (AT3G22460); ACLA-2 (AT1G60810) (Fatland et al., 2002); anthranilate synthase beta subunit (TRP4, AT1G25220) (Niyogi et al., 1993).
Table 3. Regulons represented in transcripts that are down in antisense-ACLA plants.

205 transcripts are down in antisense-ACLA plants versus WT plants. The chart shows the Regulons that are overrepresented among these 205 transcripts. For each overrepresented Regulon, the number of transcripts that are affected and the total number of transcripts are shown.

1 (Mentzen and Wurtele, 2008)

2 Fisher’s exact test was used to calculate p-values.
Table 4. MetNet pathways over-represented in down in antisense-ACLA accumulating transcripts. The down in antisense-ACLA gene list was entered into the “common pathway” tool to find over-represented pathways. The table shows the most significantly over-represented MetNet pathways occurring in the down in antisense-ACLA transcripts.

1 www.metnetdb.org

2 Fisher’s exact test was used to calculate p-values
**Figure 7. Two transcripts are up in antisense-ACLA-MA plants.** Five transcripts are up in antisense-ACLA-MA plants. Parallel coordinate plots show median gene expression data (logged, RMA normalized) for WT-H2O treated, WT-MA treated, antisense-ACLA-H2O treated, and antisense-ACLA-MA treated plants. Response regulator 16 (*ARR16*, AT2G40670, (Kiba et al., 2002)); *AtMUB5*, a ubiquitin-fold protein precursor (AT1G77870, (Downes et al., 2006)).
Figure 8. Transcription profiles of thalianol operon genes. Parallel coordinate plots showing median gene expression data (logged, RMA normalized) for WT-H₂O treated, WT-MA treated, antisense-ACLA-H₂O treated, and antisense-ACLA-MA treated plants. The thalianol operon contains four genes adjacent on chromosome 5 that together, form the thalianol biosynthesis pathway in Arabidopsis (Field and Osbourn, 2008): BADH acyltransferase (AT5G47980); THAD (thaliana-diol desaturase, AT5G47990); THAH (thalianol hydroxylase, AT5G4800); and THAS (thalianol synthase, AT5G48010).
Figure 9. GO Tree map of GO terms over-represented among the transcripts that are highest in antisense-ACLA-H₂O accumulation pattern. GO term enrichment p-values were calculated using GOstat (Beissbarth and Speed, 2004), ordered into a tree layout using GOTree (metnetdb.org), and visualized using Cytoscape (Shannon et al., 2003). The over representation of thalianol biosynthetic genes (p-value < 0.02), are represented in the triterpenoid metabolism part of the tree. Green indicates a p-value ≤ 0.05.
Figure 10. Three transcripts up in antisense-ACLA-H$_2$O. 25 transcripts in total are up in antisense-ACLA-H$_2$O plants. Parallel coordinate plots showing median gene expression data (logged, RMA normalized for WT-H$_2$O treated, WT-MA treated, antisense-ACLA-H$_2$O treated, and antisense-ACLA-MA treated plants. Expression of several genes having highest in antisense-ACLA-H$_2$O pattern, SEPALLATA 2 (SEP2, AT3G02310) (Malcomber and Kellogg, 2005), an expansin (ATEXPA10, AT1G26770) (Cho and Cosgrove, 2000), and APPETALA 3 (AP3, AT3G54340) (Malcomber and Kellogg, 2005).
Figure 11. Two transcripts down in antisense-ACLA-H₂O plants. 18 transcripts in total are down in antisense-ACLA-H₂O plants. Parallel coordinate plots showing median gene expression data (logged, RMA normalized for WT-H₂O treated, WT-MA treated, antisense-ACLA-H₂O treated, and antisense-ACLA-MA treated plants. Expression of lowest in antisense-ACLA-H₂O transcripts, a beta-ketoacyl-CoA synthase family protein (AT1G07720) and a sulfate transporter known as AST91 (AT1G23090).
Figure 12. Plots showing the accumulation pattern of metabolites. Polar and nonpolar metabolites were analyzed by GC-MS, peaks were integrated and aligned using MET-IDEA (Broeckling et al., 2006), and the resultant data were log$_2$ transformed and mean centered. Differentially accumulating metabolites (p-value < 0.05) were identified using a combination of limma (Smyth, 2005) analysis and limma/exploRase data analysis in exploRase (Lawrence et al., 2008), (Figure 4). **Panel A:** Four unidentified metabolites accumulate to highest levels in antisense-ACLA plants (a genotype effect). **Panel B:** An unidentified metabolite accumulates highest in antisense-ACLA-H2O (a genotype-by-treatment effect). **Panel C:** Serine accumulates to lowest levels in antisense-ACLA-H2O (a genotype-by-treatment effect). ACL-P, polar metabolite; ACL-NP, nonpolar metabolite; unidentified metabolites are numbered by peak elution order.
<table>
<thead>
<tr>
<th>Accumulation in ACL vs. WT</th>
<th>Model Ion</th>
<th>Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increased</td>
<td>150.2</td>
<td>Unknown</td>
</tr>
<tr>
<td></td>
<td>354.4</td>
<td>Docosanoic acid</td>
</tr>
<tr>
<td>Decreased</td>
<td>204.2</td>
<td>Glucose</td>
</tr>
<tr>
<td></td>
<td>217.2</td>
<td>Hexose</td>
</tr>
<tr>
<td></td>
<td>218.1</td>
<td>Hexose</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Accumulation in ACL-MA vs. WT-MA</th>
<th>Model Ion</th>
<th>Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decreased</td>
<td>338.2</td>
<td>Sinapinic acid</td>
</tr>
<tr>
<td>Increased</td>
<td>357.2</td>
<td>Unknown</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Accumulation in ACL-H₂O vs. WT-H₂O</th>
<th>Model Ion</th>
<th>Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Increased</td>
<td>204.2</td>
<td>Oligo-saccharide</td>
</tr>
</tbody>
</table>
Table 5. Metabolites accumulating differentially in antisense-ACLA plants and WT plants treated with malonic acid or water. Polar and nonpolar metabolites were analyzed by GC-MS; peaks were integrated and aligned using MET-IDEA (Broeckling et al., 2006). The resultant data were log2 transformed and mean centered. Differentially accumulating metabolites (p-value < 0.05) were identified using a combination of limma (Smyth, 2005) analysis and limma/exploRase data analysis in exploRase (Lawrence et al., 2008). Genotype comparisons (antisense-ACLA vs. WT) identified 5 differentially accumulating metabolites. Two differentially metabolites were identified when comparing ACLA-MA to WT-MA plants. One metabolite accumulated differentially when comparing ACLA-H2O to WT-H2O plants. Metabolite data from WT and antisense-ACLA plants was also analyzed separately to identify genotype-specific differences.
Figure 13. **Antisense-ACLA phenotype may be the result of a combination of starch and flavonoid hyper-accumulation along with a higher ratio of campesterol: sitosterol.**

Diagram showing potential interactions within the metabolome of antisense-ACLA plants. The carboxylation pathway leading to the hyper-accumulation of flavonoids is shown in purple. High levels of flavonols leads to reduced auxin transport, as indicated by the red line (Ringli et al., 2008; Santelia et al., 2008). A simplified starch biosynthetic pathway is shown in orange (Li et al., 2009). The blue condensation pathway leads to the formation of sterols, shown in black. Arabidopsis plants with suppressed SMT2 (yellow oval) activity have a high ratio of campesterol to sitosterol (Schaeffer et al., 2001), resulting in morphological characteristics (increased branched, a protruding pistil during flowering, prolonged growth, and low fertility) similar to those observed in antisense-ACLA plants (Fatland, 2002). In each mutant, the phenotype is rescued by treatment with sitosterol, but not with brassinolide (Schaeffer et al., 2001; Fatland, 2002). Red arrow shows potential decrease in the levels of sitosterol in antisense-ACLA plants, while maintaining normal levels of campesterol. Enzymes in larger type indicate those whose expression is increased in antisense-ACLA plants, with the exception of ACL itself.
CHAPTER 3: *HYPERICUM GENTIANOIDES* PRODUCES NOVEL BIOACTIVE COMPOUNDS IN SCHIZOGENOUSLY FORMED GLANDS

A paper to be submitted to the Plant Journal

Heather L Babka\(^1\), Matthew L. Hillwig\(^1\), Jason Price\(^2\), Wendy Maury\(^2\), Hilal Ilarslan\(^1\), Lankun Wu\(^1\), Eve S. Wurtele\(^1,\ast\)

\(^1\) Department of Genetics, Development, and Cell Biology, Iowa State University, Ames, IA 50011, USA
\(^2\) Department of Microbiology, University of Iowa, Iowa City, Iowa 52242, USA

Summary

The genus *Hypericum* is home to a number of plant species that have promising medicinal properties. *Hypericum gentianoides* is a North American species formerly utilized medicinally by Cherokee Native Americans. Several unusual acyl-phloroglucinols that may induce immune responses in mammalian macrophages have recently been discovered in this species. Here, we report several bioactivities of *H. gentianoides* extracts on mammalian cells and describe the ontology of secretory glands accumulating these phytochemicals in planta.

*Hypericum gentianoides* extracts alter calcium homeostasis in HEK293 cells and rat astrocytes. In addition extracts from *H. gentianoides* inhibit the infectivity of human immunodeficiency virus (HIV) HeLa37 cells in infectivity assays.

Saroaspindin A, uliginosin A, and hyperbrasilol C, the major acyl-phloroglucinols of *H. gentianoides*, accumulate within translucent glands on *H. gentianoides*. Translucent glands found on *H. gentianoides* develop schizogenously, i.e., secretory cells contained

\* For correspondence (fax +1 515 294 1337; email mash@iastate.edu).
within these glands pull away from one another to create an open cavity. Secretory cells of young translucent glands contain dense cytoplasm packed with a network of ER, ribosomes, and specialized plastids with multidirectional tubuli. Mitochondria begin to visibly degenerate in young glands and have completely disappeared by maturity. As glands age and secretory cells pull apart, bulbous protrusions appear along the cell walls, and material with a fibrillar appearance is present along secretory cell walls. Vesicles, stained dark by osmium tetroxide, occur along secretory cells walls, presumably associated with the release of secretory product into the lumen. When glands reach maturity, secretory cell size is reduced as the luminal space becomes filled with secretory product. The intensity of the osmium staining in the peripheral and surrounding cells decreases as glands age, suggestive of concomitant changes in phytochemical contents of these cells.

Introduction

Specialized phytochemicals are produced by most species of plants, providing ecologically-intricate strategies for defense against pests, diseases, or abiotic stresses, or attractants for pollination and or dispersal. Such phytochemicals are structurally diverse, and include thousands of terpenes (McGarvey and Croteau, 1995), phenolics (Bennett and Wallsgrove, 1994), methylketones (Fridman et al., 2005), phenylpropenes (Gang et al., 2001), and polyketides (Klingauf et al., 2005). Specialized phytochemicals are often sequestered into specialized cellular or subcellular compartments. For example, dhurrin, the toxic cyanogenic glucoside of Sorghum, is stored in epidermal cell vacuoles (Saunders and Conn, 1978), phenylpropenes accumulate in secretory trichomes of basil (Gang et al., 2001), α-santalene, α-bergamotene, and β-
bergamontene (sesquiterpenenes) accumulate in secretory trichomes of wild tomato
(Lycopersion hirsutum) (Besser et al., 2009; Fridman et al., 2005), nicotine and aromatic
esters (ethyl salicylate) accumulate in Nicotiana attenuata nectaries (Kessler and Baldwin,
2007), essential oil (terpenes) accumulation within oil glands of Citrus sinensis (Thomson et
al., 1976), and gum-resin accumulation in Rhus glabra secretory ducts (Fahn and Evert,
1974).

A large number of specialized phytochemicals have been exploited by humans
(Center, 2009). In particular, more than 1/3 of all pharmaceuticals are plant-derived (Li and
Vederas, 2009). For example, vinblastine and vincristine from Catharanthus roseus (Murata
et al., 2008) and taxol from Taxus brevifolia (Wani et al., 1971) have anti-cancer bioactivity,
while artemisinin from Artemisia annua is a sesquiterpenoid effective against multi-drug
resistant Plasmodium species that cause malaria and opiates (morphine, codeine) from
Papaver somniferum (Kirby, 1967).

H. perforatum, the popular herbal supplement, induces anti-inflammatory (Hammer et
al., 2007), antibacterial (Franklin et al., 2009), antiviral (Axarlis et al., 1998), and
antidepressive-like activities (Butterweck and Schmidt, 2007; Kasper et al., 2007; Kasper et
al., 2008; Mennini and Gobbi, 2004) in mammals and mammalian cell systems. The genus
Hypericum, in the Clusiaceae (mango) family, consists of over 450 different species, only a
few of which have been studied (Ernst, 2003). Interest has been fueled by the discovery that
additional species may also have medicinal properties (Ernst, 2003; Hillwig et al.,
2008; Rocha et al., 1996; Sanchez-Mateo et al., 2009; Viana et al., 2006). A few of these
species have been analyzed in more detail, and been found to accumulate unusual polyketides
including flavonoids, anthrones, xanthones, dianthrones, and acyl-phloroglucinols (Ernst, 2003; Hillwig et al., 2008; Singh and Bharate, 2006).

Several chemicals have been isolated from Hypericum species. Hypericins, which are light-activated anthraquinones, have antiviral and anticancer activity (Kirakosyan et al., 2004). Pseudohypericin is a necessary component of light-activated anti-inflammatory responses in RAW 264.7 mouse macrophage cells (Hammer et al., 2008). However, severe cytotoxic effects are associated with hypericin and pseudohypericin (Schmitt et al., 2006). These two phytochemicals cause photosensitization in unpigmented areas of skin in animals (Fields et al., 1990; Traynor et al., 2005) and death in cultured cells (Schmitt et al., 2006). Hyperforin has antimicrobial activity (Franklin et al., 2009) and is reported to be the primary bioactive component associated with the antidepressant effect of H. perforatum (Mennini and Gobbi, 2004). Dimeric acyl-phloroglucinols have been shown to have anti-bacterial properties (Rocha et al., 1996).

Plant secretory glands develop either schizogenously or lysigenously (Evert, 2006). During schizogenous development, a central lumen (open cavity) is formed as inner secretory cells pull apart; in contrast, lysigenous development entails programmed cell death and complete dissolution of the inner secretory cells (Evert, 2006). Two strikingly different kinds of internal secretory structures occur in Hypericum spp: translucent glands and black glandular nodules; it is thought that these structures serve to store some of the specialized polyketides present in Hypericum species (Curtis and Lersten, 1990). There is a paucity of sequence information for Hypericum spp, and secretory structure type(s) is a major morphological characteristic used to group different species of Hypericum into clades, (Curtis and Lersten, 1990).
Hypericum genus are those of *H. perforatum*. *H. perforatum* has both translucent glands and black nodules. The developmental processes leading to the formation of these translucent glands and black nodules appear very different, as do their molecular contents (Curtis and Lersten, 1990). Isolated translucent glands of *H. perforatum* contain hyperforin, (Soelberg *et al.*, 2007) and react positively to stains for alkaloids, lipids, resins and essential oils (Ciccarelli *et al.*, 2001b). Hypericin and pseudohypericin accumulate in the black nodules (Ernst, 2003), and ultrastructural studies suggest the secretory cells lose functionality as they fill with product (Onelli *et al.*, 2002). To our knowledge, translucent glands and canals of *Hypericum* species have not been studied at an ultrastructural level.

Several North American *Hypericum* species have been used traditionally for their medicinal properties, but have been little studied. Cherokee Native American used *H. gentianoides* (section *Brathys*, common names include orange grass and pineweed, also (inaccurately) referred to as St. Johns Wort) for the treatment of fever, gastrointestinal disorders, nose bleeds, sores, and venereal diseases (Hamel and Chiltoskey, 1975). *H. gentianoides* was reported to lack hyperforin and hypericin (Crockett *et al.*, 2005). A survey of several *Hypericum* species indicated that *H. gentianoides* produces several unusual specialized phytochemicals, in particular a range of unusual diacylphloroglucinols, including saroaspidin A, uliginosin A, and hyperbrasilol C, but no detectable hyperforin or hypericin (Hillwig, 2008).

Because of the accumulation of these unusual polyketides, the anti-inflammatory activity of *H. gentianoides* was evaluated using a RAW264.7 mouse macrophage model system (Hillwig *et al.*, 2008). In this study, the ability of *H. gentianoides* extracts to decrease accumulation of the pro-inflammatory lipid PGE₂ was measured in macrophages whose
immune response had been stimulated by lipopolysaccharide (LPS). Methanol extracts of *H. gentianoides* reduced PGE₂ accumulation in the LPS-induced macrophage cells, an indication the extract has anti-inflammatory properties; little cell toxicity was observed (Hillwig *et al.*, 2008). Iterative HPLC fractionations and bioactivity assays of *H. gentianoides* extracts indicate the molecules with anti-inflammatory activity are likely saroaspidin A, uliginosin A, and hyperbrasilol C (Hillwig *et al.*, 2008).

Here, to gain a broader idea about the potential significance of *H. gentianoides* in the ecosystem and to humans, we characterize additional bioactivities of this species on mammalian cells and evaluate the distribution and cellular location of the specialized *H. gentianoides* polyketides. We demonstrate that *H. gentianoides* extracts alter calcium homeostasis in mammalian cells, and these extracts limit human immunodeficiency virus (HIV) infectivity. Furthermore, we show that the major polyketides of *H. gentianoides*, saroaspidin A, uliginosin A, and hyperbrasilol C, are distributed in most above-ground organs of the flowering plant, and identify its discrete accumulation in translucent glands. We further investigated the ontogeny of the diacylphloroglucinol accumulating glands, showing these glands develop schizogenously and undergo major morphological and biochemical changes over time and space.

**Results**

*Growth habit of Hypericum gentianoides*

Little has been reported about the general growth and morphology of *H. gentianoides*. To gain an understanding of the general growth and morphology of this species, and to establish conditions in which it could be grown in a green house. The viability of fresh *H.*
gentianoides seeds (~90%) was found to be much higher than that of seeds after prolonged storage at room temperature (little viability after 4 weeks storage). H. gentianoides seeds were thus sown immediately after removal from mature plant seed capsules. Plants were grown in a greenhouse under natural light with a temperature of 22 – 25°C.

The development of H. gentianoides is quite different from other Hypericum species (e.g., H. perforatum, H. crux-andreae, H. androsaemum, H. hypercoides, H. deticulatum) in that H. gentianoides has a distinctive vegetative and flowering stage (Figure 1 a - b). Diversity is also evident among flowers, which are significantly smaller than that of other Hypericum (Figure 2). In the vegetative stage, H. gentianoides has a compact growth stage, with shortened internodes, approximately 5 cm tall and normal leaves (Figure 1 a). As H. gentianoides starts flowering, the internodes rapidly lengthen and the plant quickly becomes tall, spindly, and covered with leaves modified into appressed, subulate scales, and this process is referred to as bolting (Figure 1 b). Flowers are present at each node on an inflorescence (Figure 1 a). The tip of each inflorescence usually has one large flower bud and a smaller, younger flower bud. The internodes contain two small leaves, with the youngest leaves at the internode closest to the floral tip. Visual inspection of H. gentianoides reveals translucent glands on almost all above-ground tissues. No other types of secretory-like structures are detectable on any part of the plant. Stereomicroscopic and visual observations indicate the largest number of glands occur on leaves and flower buds with some glands also present on stems (Figure 1 e). No glands are detectable on flower petals or the gynecium. Removal of pigments and other material from leaves by emersion in a series of solvents (‘clearing’, (Ilarslan et al., 2001), reveals that these glands are located in the regions between veins (Figure 1 c – d).
Hypericum gentianoides methanol extracts exhibit bioactive properties when applied to mammalian cells

Hypericum gentianoides extracts induce transient cytotoxic Ca\(^{2+}\) increase in mammalian cells

Fluctuation in calcium ion concentration provides a key early signaling mechanism in virtually all eukaryotic cells (Berridge et al., 1998; Luan, 2009) and plays a major role mediating responses of the immune system (Feske, 2007), neuronal cells (Patterson et al., 2007), and smooth muscle cells (Leung et al., 2008). To better understand the early events that might impact the anti-inflammatory response induced by *H. gentianoides* extracts (Hillwig et al., 2008), we evaluated the effect of *H. gentianoides* extracts on calcium flux in mammalian cells. In these studies, extracts were made from above-ground material of approximately 5 month-old, *H. gentianoides* plants (as in Figure 1 a) and applied to two diverse cell types: cultured HEK293 (human embryonic kidney) cells (Shaw et al., 2002) and primary cultures of rat astrocytes (Perea and Araque, 2005). The HEK293 cell line was chosen because its transcriptome and physiology are well-characterized (Shaw et al., 2002) and it is widely used as a transfection model for receptors (Kumpost et al., 2008), thus is a good model for future studies. Astrocytes were selected as a second cell line for study because they are thought to play an important role in immune response (Dong and Benveniste, 2001; Miljkovic et al., 2007) and H. gentianoides has already been implicated in this response (Hillwig et al., 2008).

Cytosolic Ca\(^{2+}\) concentrations ([Ca\(^{2+}\) ]\(_i\)) were monitored by radiometric imaging of individual cell traces over time using Furan 2-dye as an indicator of [Ca\(^{2+}\)]\(_i\) (Patterson et al., 2007). Data are compiled from cell traces of a minimum of 60 independent cells, obtained
during three separate experiments, each using a different batch of cultured cells (i.e., traces of over 20 cells/biological replicate).

\textit{H. gentianoides} extracts evoke a transient cytosolic calcium increase in HEK293 cells (Figure 3a and c). This \textit{H. gentianoides}-evoked calcium response returns to baseline after removal of \textit{H. gentianoides} extract, indicating that the activation is reversible. At a dose of 50 μg extract/ml, the increase of [Ca$^{2+}$] was 100.23 ± 20.15 nM (Figure 3 d).

A rat astrocyte primary cell culture model was selected to further test the activity of \textit{H. gentianoides} extracts. The astrocyte is the most common cell type in the central nervous system (CNS), constituting 50% of the total CNS cells in the adult (Song \textit{et al.}, 2002). Under study of astrocyte biology has recently expanded; they are important in synapse formation (Slezak and Pfrieger, 2003), capable of releasing neurotransmitters (Newman, 2003), and, key to the current study, they play a role in the immune response and in cytokine production (Dong and Benveniste, 2001; Miljkovic \textit{et al.}, 2007). \textit{H. gentianoides} extracts induce transient cytosolic calcium increase in primary cultured rat astrocytes. At a dose of 50 μg extract/ml, [Ca$^{2+}$]$_i$ was 46.28 ± 7.41 nM (Figure 3 d). This is the first report of bioactivity of any \textit{Hypericum} species in cells from the CNS.

\textit{Hypericum gentianoides} extracts reduce the infectivity of HIV

\textit{H. perforatum} has inhibitory effects on the infectivity of lentiviruses, including HIV, in mammalian cells (Richman, 1991), but also causes undesirable effects. These include hypericin-induced photodermatitis (Traynor \textit{et al.}, 2005) and hyperforin-induced up-regulation of cytochromes, which increase the metabolism of co-administered drugs (Gödtel-Armbrust \textit{et al.}, 2007). Hyperbrasilol B, which accumulates in \textit{H. connatum} and is used in
Brazil for treating oral sores, has been shown to reduce the viral levels of herpes simplex virus-1 (Fritz et al., 2007). Because *H. gentianoides* extracts contain related polyketides (including the isomer hyperbrasilol C) and have little cytotoxicity against macrophages (Hillwig et al., 2008), we evaluated extracts for their potential to influence the infectivity of HIV in a standard HeLa37 cell assay (Reed-Inderbitzin and Maury, 2003).

At a concentration of 10 μg/ml, methanolic extracts of *H. gentianoides* reduce HIV infectivity by up to 50%, with no detectable change in cell viability (Figure 4 a). At higher doses (100 μg/ml) extracts are highly cytotoxic, resulting in almost total death of the HeLa37 cells. To further evaluate this effect, methanol extracts were prepared from three batches of material of two different accessions of *H. gentianoides* (Ames 28015 and Ames 27480) and these extracts were analyzed for their ability to inhibit HIV infectivity of HeLa 37 cells (Figure 4 b). One extract (Ames 27480 - Fresh) was cytotoxic to the HeLa37 cells even at doses of 3 μg/ml (Figure 4 b). In contrast, extracts from dried Ames 28015 reduced HIV infectivity at an application of 3 μg/ml by about 20%, and at a 10 μg/ml application reduced HIV infectivity by 40%, with no significant effect on cell viability (Figure 4 b). A second extract from Ames 28015 – dry reduces the ability of HIV to infect HeLa37 cells at 10-30 μg/ml with little cytotoxicity. Thus, the ability of *H. gentianoides* to impact HIV infection and or cytotoxicity is highly dependent on factors such as the dosage, the particular accession and/or the method of storage of plant material.

*The bioactive compounds, saroaspidin A, uliginosin A and hyperbrasilol C preferentially accumulate in flower buds and young leaves of Hypericum gentianoides*
In order to determine which plant organs and stages produce the highest amounts of
diaclyphloroglucinols, flower buds, nodes (plus subtending leaves), and internodes (Figure 5 b) from ten inflorescences of *H. gentianoides* were dissected, extracted with methanol, and fractionated by HPLC. Saroaspidin A accumulates significantly in nodes 2 and 3, while hyperbrasilol C and uliginosin A accumulate to high levels in the nodes and flower buds (Figure 5 a). The internodes are not enriched with any of these compounds.

_Gland contain saroaspidin A, uliginosin A and hyperbrasilol C_

In order to isolate glandular contents a capillary needle method was developed in which capillary tubes were partially melted in the flame of a Bunsen burner, stretched out by pulling opposite ends away from one another, and broken manually to make small-diameter tubes (~ 50 μm) that could puncture *H. gentianoides* glands. The resultant tubes were inserted into glands from multiple leaves (250 glands total), and gland contents traveled up the tube via capillary action. Glandular material was removed by forcing air into the tube and dissolved in 100% methanol (500 μL). The resultant extract was directly injected into an electrospray ionization (ESI) mass spectrophotometer.

Mass spectral data indicates glands contain primarily the three major anti-acyl-phloroglucinols of *H. gentianoides*: saroaspidin A, uliginosin A and hyperbrasilol C (Figure 6). Furthermore, the ratios of saroaspidin A: uliginosin A: hyperbrasilol C in isolated glands (Figure 6) is similar to their ratio in extracts from the second node plus its subtending leaves (Figure 5).
Morphology and ultrastructure of the translucent glands

Previously, translucent glands of *Hypericum* have not been visualized using electron microscopy, and virtually nothing is known about the glands of *H. gentianoides*. To define the ontogeny of the glandular structures containing saroaspidin A, uliginosin A and hyperbrasilol C, plant organs were subjected to light microscopy (LM), scanning electron microscopy (SEM), and transmission electron microscopy (TEM). Since the acyl-phloroglucinols accumulate to high levels in both flower buds and young leaves (Figure 5 b), and these are plentiful, these organs were the focus of analysis.

Visual observations of dissected fresh plant material indicate flower buds of *H. gentianoides* contain large numbers of glands in the outer protective leaves and sepals; in contrast, no glands are visible in the petals. To examine the glandular distribution and structure on flower organs in more detail, cross and longitudinal sections of flower buds were examined by LM (Figure 7 a - b). Longitudinal sections of flower buds show glands of different developmental stages are located on sepals surrounding the floral meristem (Figure 7 a), with younger glands tending to be located near the tips and base of sepals; petals do not have glands. A single large gland is located in the tip of each anther (Figure 7 c - d). This central gland is embedded in the connective tissue at the anther’s apical end, and is surrounded by calcium oxalate crystals (Figure 7 c – d). Crystals are prominent within cleared, whole *H. gentianoides* anthers (Figure 7 e). Individual crystals were isolated using a surgical blade and dissecting needle for LM (Figure 7 f) and SEM analysis (Figure 7 g– j) in order to observe the crystal structure. The crystals isolated from *H. gentianoides* anthers are large spherical druses, composed of multiple crystal faces. SEM images indicate the anther gland contains a putative secretory substance (Figure 7 j).
Based on the large number and size of glands on the sepal, and the apparent
developmental gradient of stages on each sepal, we focused on sepals to investigate glandular
development. Detailed microscopic observations from approximately 15 glands on 20 sepals
from flower buds of different ages revealed a pattern in gland development. We were able
to assign glands to developmental stages based on the following criteria: Stage 1 glands
contain cytoplasmically dense secretory cells, small, uniform peripheral cells, and little to no
luminal cavity (Figure 8 a); Stage 2 glands have cytoplasmically dense secretory cells,
osmium-stained peripheral cells, degraded cell wall and the beginning appearance of a lumen
(Figure 8 b); Secretory cells of stage 3 glands are much reduced and remain cytoplasmically
dense, peripheral cells are highly osmium-stained, and an enlarged luminal cavity is present,
sometimes observed containing an osmium-stained droplet (Figure 8 c); By stage 4,
secretory cells are becoming filled with vesicle-like structures, peripheral cells stain less
intensely with osmium, and a lumen is present (Figure 8 d); Stage 5 gland secretory cells
contain enlarging vesicle-like structures and are less cytoplasmically dense, peripheral cells
are less densely osmium-stained, and lumen is present (Figure 8 e); Stage 6 glands have a
greatly reduced secretory cell volume as the cells become appressed to the peripheral cells,
peripheral cells are highly vacuolated, and the gland is mostly lumen (Figure 8 f).

Young glands (stages 1 and 2)

Young translucent glands of *H. gentianoides* are packed with cytoplasmically dense cells
(Figure 8 a – b). The dense cytoplasm of these secretory cells can be seen in detail using
TEM (Figure 9 a). Within the cytoplasm of stage 1 and 2 gland secretory cells is a dense
network of ER and associated ribosomes (Figure 9 d, e); the abundance of ribosomes may
indicate active translation, perhaps to synthesize the enzymes necessary for secretory product synthesis. Unusual plastids are apparent; these are elongated and filled with multidirectional tubuli (Figure 9 c - e), such plastids containing multidirectional tubuli are common in secretory duct cells (Fahn, 1979). “Ghost-like” presumably degenerating mitochondria are present, and “normal” appearing mitochondria were not observed (Figure 9 b – d), possibly indicating that these young secretory cells no longer require the energy or metabolic products from mitochondria. The secretory cells within stage 2 glands contain small bodies that stain with osmium (Figure 9 d - f).

The outer cell walls (separating the outer peripheral cells and inner secretory cells) are thicker than the walls separating secretory cells in both stage 1 and 2 glands (Figure 9 b). Vesicle-like structures line the secretory cell walls in a zipper-like fashion (Figure 9 b). Cell wall breakdown, which becomes apparent by stage 2, is characterized by bulbous growths along the secretory cell walls and breakdown of the middle lamella (Figure 9 c).

The peripheral cells surrounding stage 1 and 2 glands contain chloroplasts with starch granules and grana as well as normal-appearing mitochondria (Figure 9 b, f). Vacuoles of peripheral cells are large and stain heavily with osmium. The cytoplasm of peripheral cells is significantly less dense that of the secretory cells. All these are characteristics of a metabolically active cell.

_Developing glands (stages 3 and 4)_

Developing glands are characterized by the increasingly thin line of secretory cells pushed against the peripheral cells (Figure 8 c - d, 10 a). Figure 8 c shows a droplet of lightly osmium-stained material located in the lumen of a stage 3 gland. The major distinguishing
factor separating stage 3 from stage 4 glands is the greater cytoplasmic density of stage 3 secretory cells. The secretory cells of stage 3 and 4 glands contain more vesicle-like structures compared to earlier stages, some of which are stained darkly with osmium (Figure 10 b – i). In general the vesicle-like structures in stage 3 secretory cells appears to be more densely stained with osmium, while at stage 4 the osmium-stained material within the vesicle-like bodies is more granular.

Specialized plastids containing multidirectional tubuli, ER, and ribosomes are present in the cytoplasm of stage 3 secretory cells, while mitochondria are no longer obvious (Figure 10 e – f). These plastids do not appear to be the source of the osmium-stained vesicle-like structures found in the secretory cells, as no intermediate structures are observed.

The cell walls of both stage 3 and 4 secretory cells have become very thin compared to the walls of the outer peripheral cells (Figure 10 c – e, g, i, k). Secretory cell walls are very deformed, with many bulbous areas on the luminal side of the cell (Figure 10 b, k). The fibrillar projections observed along secretory cell walls on the lumen side could be pieces of degraded cell wall and/or secretory material. Vesicle protrusions containing an osmiphillic substance appear along secretory cell walls (Figure 10 b & k). The lumen of the gland contains material lightly stained with osmium, which is likely the secretory substance.

By stage 4, the chloroplasts of the peripheral cells still contain starch, but they are starting to look degenerated (Figure 10 k). Peripheral cells (stage 4) have reduced cytoplasmic density as compared to those found in stage 1 and 2 glands (Figure 10 g, k). By this time the amount of osmium-stained material present within peripheral cell vacuoles is reduced (Figure 10 a).
**Mature Glands (stages 5 and 6)**

In mature glands (stages 5 AND 6) the secretory cells are almost completely free of cytoplasm as the secretory phase of these cells ceases (Figure 11). Specialized plastids can still be observed in the cytoplasm of stage 5 secretory cells, but in much reduced numbers (Figure 11 b, f). These were not observed in stage 6 glands. By stage 6, dark osmium-stained droplets appear contained within enlarged vacuoles. (Figure 11 g – k)). The large lumen of the mature gland contains lightly-stained secretory material (Figure 11 a - b, d, f, k). Along the cell walls of secretory cells in mature glands material appears to be sloughed off the walls into the central lumen (Figure 11 i - j). Again, it is unknown whether this fibrillar material is cell wall debris or secretory product deposited via direct diffusion.

The vacuoles of the peripheral cells are also enlarged (stage 5 and 6) and chloroplasts are still present in the cytoplasm of stage 5 glands (Figure 11 a & c). By stage 6 peripheral cells appear to be filled with primarily particulate material (Figure 11 h).

The ontogenesis of leaf glands was similarly examined by TEM; it is indistinguishable from the development of sepal glands (data not shown).

**Discussion**

Plants, being rooted in place, often make secondary compounds as a defensive mechanism. Some of the specialized phytochemicals that may act as anti-feedants are also highly toxic to plant cells, and the phytochemicals are sequestered in specialized structures (Ernst, 2003; Gershenson et al., 2000; Turner et al., 1998) The products of these structures are released when an unsuspecting herbivore and/or insect takes a bite of the plant (Fahn, 2002; Pickard, 2008). The presence of large numbers of glands in young tissues, especially
around meristematic areas, indicates that the glands of \textit{H. gentianoides} could serve to deter herbivores.

The results presented here demonstrate that saroaspidin A, uliginosin A, and hyperbrasilol C, the major diacylphloroglucinols of \textit{H. gentianoides}, accumulate to high levels in abundant translucent glands covering the leaves and sepals. Schizogenous development is one common developmental pattern for secretory ducts, for example, gum resin ducts in \textit{R. glabra} (Fahn and Evert, 1974) and oil glands on \textit{C. sinensis} (Fahn, 1979). To date this is the first study of the development of translucent glands at the ultrastructural level in order to understand the developmental process.

Translucent glands and secretory canals of \textit{H. perforatum} can be further categorized into three different classes (referred to as type A, type B, and type C) differing by location (on plant) and morphology (Ciccarelli \textit{et al.}, 2001b). Type A secretory canals, found in all floral tissues and in stem and root cortical parenchyma, are narrow and delimited by four polygonal cells and undergo schizogenous development. Type B secretory canals are quite large, resembling elongated translucent glands and are located in sepals and petals and develop schizogenously. Type C secretory canals, are located in the ovary and style, have a morphology similar to type B secretory canals, but develop, using a combination of schizogenous and lysigenous processes (Ciccarelli \textit{et al.}, 2001b). We did not observe type A or type C canals in \textit{H. gentianoides}, while elongated type B secretory canals can be found on sepals.

The black nodular glands of \textit{H. perforatum} are distinct from secretory structures observed in other genera (Ciccarelli \textit{et al.}, 2001a; Curtis and Lersten, 1990). Unlike translucent glands, at nodule maturity, the cells within the nodule are completely filled with
hypericin and other compounds (tannins, alkaloids) (Ciccarelli et al., 2001a; Curtis and Lersten, 1990). Ultrastructural studies show that these secretory structures lack a lumen and presumably the secretory cells lose functionality and become storage places (Onelli et al., 2002). When black nodular glands of two different H. perforatum shoot cultures were compared morphologically, the glands of the line producing higher amounts of hypericin and psuedohypericin had larger peripheral cells (Kornfeld et al., 2007).

Little is known about the biosynthesis of the specialized polyketides in Hypericum. Polyketide synthases (PKS) catalyze sequential reactions involving decarboxylative condensation of carboxylic acid units (Gokhale et al., 2007). The modularity of PKSs allows for the production of a wide variety of molecules (ref). This inherent complexity results in a remarkably wide range of bioactivities (i.e., anti-inflammatory, anti-viral, anti-microbial). Type III PKSs are thought to have evolved from the fatty acid biosynthetic enzyme, keto acyl-synthase III (KASIII) (Austin and Noel, 2003). Hyperforin is synthesized from isobutyryl-CoA by the Type III PKS butyrophenone synthase (BUS), in H. calycinum (Klingauf et al., 2005). The expression of the Type III PKS, HpPKS2, within dark nodules containing hypericin in H. perforatum has been reported and is thought to catalyze hypericin synthesis in this species (Karppinen et al., 2008).

Hypericum gentianoides glands develop schizogenously

The schizogenous development observed in H. gentianoides translucent glands, is consistent with light microscopic observations made about H. perforatum translucent gland development (Ciccarelli et al., 2001b; Curtis and Lersten, 1990). Thus it seems likely that
more detailed ultrastructural observations made for *H. gentianoides* translucent glands will apply to *H. perforatum* translucent glands.

Young *H. gentianoides* glands are packed with cytoplasmically dense secretory cells and surrounded by biologically-active peripheral cells. Secretory cells in young glands appear highly active, containing a dense network of ER, large numbers of ribosomes, and plastids filled with multidirectional tubuli. Multidirectional tubuli have also been observed in plastids from *Poncirus trifolata* secretory glands (Fahn, 1979). Mitochondria in *H. gentianoides* secretory cells appear degenerated, even in young glands (stage 1), this is very different than observations from *R. glabra* (Fahn and Evert, 1974) and *Dahlstedtia pentaphylla* (Teixeira and Rocha, 2009) secretory cells in which mitochondria are “healthy” looking. In general, secretory cells contain many mitochondria, presumably supplying ATP and substrate to support active secretory product synthesis (Fahn, 1979). In *H. gentianoides* secretory cells, either the mitochondria do not play an important role in the synthesis of the secretory product or the usefulness of mitochondria has ended in the stages of development observed in this study. The latter is the most reasonable explanation, but it is surprising that mitochondria do not persist into stages 4 – 6.

The major difference between stage 1 and stage 2 glands is the degenerated appearance of secretory cell walls in stage 2 which likely represents the beginning of the secretory process (Figure 9). Prior to secretory cell wall thinning, vesicle-like structures were observed along secretory cell walls adjacent to the lumen which possibly contain cell wall degrading enzymes, because later secretory cell walls appear bloated and the middle lamella is less structured (Figure 9 c). LM observations of *H. perforatum* translucent glands also have shown vesicle-like structures near the secretory cell walls (Ciccarelli *et al.*, 2001b); similar
structures thought to be filled with cell wall digesting enzymes, such as pectinases have been observed in C. sinensis (Thomson et al., 1976) and R. glabra (Fahn and Evert, 1974) prior to wall breakdown. Beginning at stage 3, small, vesicle-like structures containing lightly osmium-stained material can be found along degraded secretory cell walls and they are no longer obvious by stage 5. These vesicle-like structures likely contain secretory product.

The fibrillar material, located on the lumen side of the secretory cell wall, could be cell wall debris or even secretory product.

The reduction in the size of secretory cells in stages 5 and 6 (Figure 11) probably reflects a reduction in secretory cell activity combined with increased luminal content. Fewer organelles are present and cytoplasmic density is much reduced in secretory and peripheral cells in stages 5 – 6, as the activity in both cell types’ winds down. Secretory product is found within the lumen, while darkly staining droplets are present in enlarged secretory cell vacuoles at stages 5 and 6 (Figure 11). The dark, osmium–stained droplets are pronounced in the vacuoles of stage 6 secretory cells. It is likely that the smaller dark vesicles-like structures found in stages 3 and 4 (Figure 10 h) fuse to make the larger oil bodies observed in stages 5 and 6 secretory cells. The osmium-stained droplets are being sequestered within the vacuoles of secretory cells to separate the presumably phytotoxic compounds from important cellular processes.

**Spatiotemporal accumulation of dark osmium-stained material**

This study reveals saroaspidin A, uliginosin A, and hyperbrasilol C accumulate within glands; however the compartmentalization of the biosynthetic process is not known. Indeed compartmentalization has not been described for the biosynthesis of any other polyketides in
plants. In the few instances where the spatial localization of a chemically complex metabolite has been described such as, vindoline biosynthesis being localized within multiple tissue types in *C. roseus* (Murata *et al.*, 2008) and the involvement of multiple subcellular compartments in isoprenoid biosynthesis in Arabidopsis (Phillips *et al.*, 2008), the biosynthetic process occurs across time and space.

It is likely that saroaspisin A, uliginosin A, and hyperbrasilol C are synthesized within the secretory cells and peripheral cells/surrounding cells. During development of *H. gentianoides* glands, the composition and/or concentration of the material in the cells surrounding the gland changes. Early in gland development, the peripheral cells of the gland, as well as some cells surrounding the gland, stain darkly with osmium. Over time, the intensity of the osmium stain changes. The content of the peripheral cells begins to stain less darkly with osmium beginning around stage 5, indicating that the properties or concentrations of chemicals has changed. This phenomenon has been observed in the maturation of oil glands of *C. sinensis* (Thomson *et al.*, 1976). In *H. perforatum* black glandular nodules the size of peripheral cells has been associated with the amount of hypericin produced, indicating peripheral cells may be involved in the production of hypericin (Kornfeld *et al.*, 2007). Thus, it would not be farfetched for the peripheral cells of *H. gentianoides* to have some involvement in the biosynthesis of saroaspisin A, uliginosin A, and hyperbrasilol C.

We have no indications as to how the precursors get into secretory cells. Structures that are specialized in massive transport of small molecules, such as transfer cell (Offler *et al.*, 2003) or plasmodesmata (Fahn and Evert, 1974) were not observed in *H. gentianoides* glands. We suspect these data indicate that the darkly staining peripheral cells contain acyl-
phloroglucinol precursors that are later utilized by secretory cells in the production of saroaspidin A, uliginosin A and hyperbrasilol C.

In order to determine the contents of secretory cells and the surrounding tissue, an in situ approach is needed. Graphite-assisted laser desorption ionization MS (GALDI-MS) has been used to profile the accumulation location of flavonoids in the flowers and sepals of Arabidopsis (Cha et al., 2008); as the resolution of this technology improves it could potentially be used to determine the contents of peripheral cells, secretory cells, and surrounding tissues of glands in *H. gentianoides*.

*Accumulation of saroaspidin A, uliginosin A, and hyperbrasilol C*

Because of the very high concentration of the three phloroglucinols in the glands and because the ratios of saroaspidin A: uliginosin A: hyperbrasilol C within tissues containing large numbers of glands and the isolated glands are similar, we predict the glands are the primary site for the storage of these compounds. Certainly the secretory cells are involved in the deposition of secretory material into the luminal cavity of the gland. The secretory material probably enters the lumen either via diffusion across the cell wall or through deposition by vesicles (Fahn and Evert, 1974). Beginning at stage 3, material accumulates within the lumen of the gland, including fibrillar material and lightly osmium-stained vesicle-like structures. The light osmium-stained vesicle-like bodies persist into stage 4, while fibrillar protrusions remain until stage 6. The presence of fibrillar material could indicate that direct or active diffusion of product (ecrine secretion) into the lumen occurs, and that this process continues late into gland development. On the other hand, the presence of vesicle-like structures filled with lightly osmium-stained material, indicates granulocrine secretion, which
ends much earlier in gland development. It is unclear from our observations which method is primarily responsible for deposition of secretory material within the lumen.

The anther’s secretory gland.

Visual inspection revealed a single, large translucent gland on the anthers of *H. gentianoides*. In other *Hypericum* species, the anther filament terminates in a dark or amber colored gland (Ernst, 2003). The *H. gentianoides* anther gland is surrounded by calcium oxalate crystals; such crystals have been identified in anther connective tissue of several other species including sunflower (*Heliantheae*) (Schmid, 1976), *Lycopersicon esculentum* (Bonner and Dickinson, 1989), *Petunia* (Iwano et al., 2004) and *Capsicum annuum* (Horner and Wagner, 1992). The role of these crystals is not well understood. Calcium oxalate crystals may play a role in calcium storage (Nakata, 2003). *C. annuum* anthers contain four different types of calcium oxalate crystals, suggesting differing transport of micronutrients in cells (Horner and Wagner, 1992). Crystal formation in *C. annuum* (Horner and Wagner, 1992) and *L. esculentum* (Bonner and Dickinson, 1989) anther stomium tissue has been implicated in anther dehiscence. It would not be surprising if the calcium oxalate druses found in *H. gentianoides* anthers play a similar role in pollen maturation and release.

Bioactivity of acyl-phloroglucinols from *Hypericum gentianoides*

Methanolic extracts from *H. gentianoides* reduce PGE$_2$ concentrations in LPS-induced macrophages (Hillwig et al., 2008). Although this is also true also for *H. perforatum* (Hammer et al., 2007), the metabolite profiles differ greatly between *H. perforatum* and *H. gentianoides* (Hillwig, 2008). In fact, AFLP markers indicate that these two species are only very distantly related to one another within the 400-plus membered *Hypericum* genus.
(Percifield et al., 2007), yet each sequesters distinct compounds in specialized glands and these compounds ultimately have anti-inflammatory properties, but very different cytotoxicity (Hammer et al., 2007; Hillwig et al., 2008). There is much to be gained by understanding if/how these different compounds might be acting on targets in mammalian cells. The sequestering of phloroglucinols in the “clear” glands of Hypericum species suggests they may be toxic to the plant. One plausible effect that might explain the sequestering of these compounds is that they may interfere with calcium signaling as they do in mammalian cells.

Related to the chemical ecology, herbivory could release the contents of the glands, disrupting the grazer’s digestive tract, and also cause signaling activation within the plant itself by binding to certain receptors. This is not entirely implausible, considering there are numerous volatiles present in Hypericum, particularly monoterpenes that could diffuse rapidly in the air and signal a defense response in nearby plants. Indeed, the acyl-phloroglucinols could also act as anti-oxidant “preservatives” for the terpenes, since phloroglucinols have anti-oxidant properties; such anti-oxidant activity could also explain some of the disease-fighting properties of this class of phytochemicals (Verotta, 2002). In addition, acyl-phloroglucinols can be chemically modified, for example they could be prenylated; this would increase their solubility in very non-polar environments such as the glands. It could also result in destruction of these compounds into volatile signaling molecules; such volatiles would likely not be detected by our LC-MS-MS or GC-MS protocols.

Translucent glands of H. gentianoides contain novel bioactive acyl-phloroglucinols: saroaspidin A, uliginosin A, and hyperbrasilol C. H. gentianoides methanolic extracts, having
high levels of these specialized phytochemicals (Hillwig, et al., 2008), cause transient increases to cytosolic calcium levels in HEK293 cells and primary rat astrocytes. The ability of \textit{H. gentianoides} extracts to increase intracellular calcium levels in very diverse cell types indicates that the bioactivity may be modulated by more than one type of receptor. We postulate that endogenous \textit{H. gentianoides} compounds, likely acyl-phloroglucinols, activate specific receptors in HEK293 cells and astrocytes. In fact, preliminary data suggests that isolated uliginosin A and hyperbrasilol C both have bioactivity in evoking a transient cytosolic calcium increase in HEK293 cells and rat astrocytes (Wu, Hillwig, and Wurtele, data not shown). We have not yet identified which receptor(s) these compounds are interacting with. However, this work will help to set the stage for evaluation the potential mechanism(s) of the reported anti-inflammatory activity of this plant (Hillwig et al., 2008).

Extracts of \textit{H. gentianoides} were also able to limit the infectivity of HIV in HeLa37 cells. A number of acyl-phloroglucinols have with biological activities including anti-viral effects (Singh and Bharate, 2006). This result, combined with the ability of \textit{H. gentianoides} extracts to affect the calcium status of two distinct mammalian cell types, make further study of the modes of action important.

**Summary**

PKS enzymes synthesize a wide variety of economically important specialized phytochemicals having a wide range of bioactivities. Currently, only three Type III PKSs from \textit{Hypericum} species have been described: BUS from \textit{H. calycinum} (Klingauf et al., 2005), BPS from \textit{H. androsaemum} (Benye et al., 2003), and HpPKS2 from \textit{H. perforatum} (Karppinen et al., 2008). A technique such as laser capture microdissection (LCM) could be
utilized in *H. gentianoides* in order isolate glandular areas to create a cDNA library enriched for the enzymes responsible for the biosynthesis of saroaspidin A, uliginosin A, and hyperbrasilol C. Isolating the PKS enzymes (and genes) leading to the biosynthesis of saroaspidin A, uliginosin A, and hyperbrasilol C is important for having a better understanding of the modularity these enzymes and their ability to make a plethora of important compounds. It is also important to be able to compare how the enzymes that synthesize saroaspidin A, uliginosin A, and hyperbrasilol C differ and how these differences influence the chemistry of the compound produced. Having a better understanding of the enzymes responsible for the biosynthesis of saroaspidin A, uliginosin A, and hyperbrasilol C as well as isolation of the genes encoding the enzymes, will facilitate the creation of a more cost-effective therapy, should one ever be developed from *H. gentianoides*.

Here, we report two potentially significant bioactivities of *H. gentianoides*, and detail the *in planta* distribution of its major specialized phytochemicals. Having revealed the glandular location of saroaspidin A, uliginosin A, and hyperbrasilol C accumulation, and established the general processes involved in the development of the *H. gentianoides* glands, sets the stage for more directed studies to determine the cellular site(s) and mechanisms of transport of the saroaspidin A, uliginosin A, and hyperbrasilol C and their precursors, and to isolate the genes and enzymes responsible for the biosynthesis of these acyl-phloroglucinols. Understanding the pathway and enzymology for the production of biologically active phytochemicals, as well as the associated sequestration of this process, is key to making these, as well as potentially bioactive derivative compounds, widely available as treatments.
Experimental Procedures

Plant materials and growth

*H. gentianoides* germplasm, accessioned in the US National Plant Germplasm System as Ames 27629, Ames 27657 and Ames 28015 (voucher specimens located at Iowa State University Ada Hayden Herbarium, www.public.iastate.edu/~herbarium/), was obtained from the USDA North Regional Plant Introduction Station, Ames, IA. Information about Ames 27629 is posted in the Germplasm Resources Information Network (GRIN) database at http://www.ars-grin.gov/npgs/.

*H. gentianoides* was grown in a soil mixture (65% peat moss, 30% perlite, and 5% mineral soil) in a greenhouse with natural light at 22 – 25°C. Non-flowering *H. gentianoides* was 3 – 4 months old at harvest, while flowering plants were 5 – 6 months of age.

Metabolite analysis of extracts

The bud, first three nodes plus subtending leaves, and internodes were dissected from bolting *H. gentianoides* using two biological replicates (ten stems each) and frozen in liquid nitrogen. Plant material was ground in a liquid nitrogen-cooled mortar and pestle to a fine powder. Then 1.5 ml of methanol was added to the mortar and pestle and mixture was ground for an additional 30 sec. The resultant extract was collected into Eppendorf tubes, incubated at 70°C for 2 min to deactivate enzymes, and immediately re-frozen in liquid nitrogen. The extracts were centrifuged at 12,000 rfc for 2 min and the methanolic supernatant was removed and filtered through a 2 μM nylon syringe filter. The resultant methanolic extracts were analyzed using a Beckman Coulter high performance liquid
chromatography instrument with a photodiode array Detector 160, and peaks were normalized by sample weight (Hillwig, 2008; Hillwig et al., 2008). The peaks included three bioactive acyl-phloroglucinols, three suspect acyl-phloroglucinols, and a suspected sterol as determined by Hillwig (2008).

**Metabolite analysis of glands**

Capillary tubes were held in the flame of a Bunsen burner, pulled, and broken to obtain pulled tubes with a small diameter (small enough to only puncture a gland). Using pulled needles and a stereomicroscope, the contents of 250 individual glands of *H. gentianoides* vegetative leaf tissue were collected via capillary action. The pulled-capillary needle technique was too difficult to perform on areas of the leaf without glands, since these areas contain major vascular strands that break the needles. The collected liquid was expelled and dissolved in 500 µl of methanol. This extract was directly infused into an Agilent Ion Trap 1100 electrospray-ionization mass spectrometer at 2 µl/min.

**Semi-preparative HPLC**

A Synergi Max-RP 4 micron 250x10 mm column (Phenomenex Torrance, CA 90501) was used on a Beckman Coulter HPLC with a Detector 160 PDA detector. For the mobile phase, an acetonitrile/methanol 9:1 v/v (solvent B) and 10 mM ammonium acetate (solvent A) gradient elution was used; 13-15% B in 10 min, 15-100% B in 30 min, 100% B for 5 min, at 40°C. All solvents used were HPLC grade (Sigma, St. Louis, MO). The extract was injected repeatedly for fractions to yield enough material for assays (Hillwig, 2008). All fractions collected were dried under nitrogen gas and re-solubilized in cell culture grade DMSO for cell assays (Sigma, St. Louis, MO).
**HEK293 cell culture**

HEK293 cells were obtained from ATCC (Manassas, VA, USA) and cultured in Dulbecco’s modified Eagle’s Medium (DMEM) supplemented with 10% fetal bovine serum (FBS), 50 U/ml penicillin, 50 μg/ml streptomycin, and 2 mM glutamine. Cells were grown in an incubator at 37°C with humidified 5% CO₂ and 95% air (Veterinary Medicine, Iowa State University).

**Astrocyte cell culture**

Mixed neuron glia cultures (Veterinary Medicine, Iowa State University) from freshly dissected neonatal rat hippocampus were established as described (Slezak and Pfrieger, 2003). Tissue was treated in 2 ml of papain solution (1.54 mg/ml of Earle’s Balanced Salt Solution (EBSS), 40 min at 37°C), then washed with EBSS, incubated 5 min in trypsin inhibitor, and mechanically dissociated in DMEM high glucose + 10% FBS, + penicillin / streptomycin (1 mL/100 ml DMEM) using pipettes. Cells were plated in DMEM medium in culture flasks and maintained at 37°C in a humidified 5% CO₂/95% air atmosphere. After mixed cultures reached confluence (9-12 days), the flasks were "preshaken" (260 rpm) for 90 min to remove microglia and dividing type I astroglia. The cultures were then shaken overnight (12-18 h) at 260 rpm at 37°C. Cultures enriched in type I astroglia were obtained by trypsinizing (0.25%) the attached cells for 3 min. Trypsin was inactivated by adding 3 ml DMEM (the same media as above) supplemented with 10% heat-inactivated FBS (serum contains protease inhibitors). Astrocytes were plated on poly-L-lysine (100 mg/ml; MW 100,000)-coated glass coverslips. All experiments were performed on cells in culture for 1-3 days after transfer to new media.
**Intracellular Calcium Imaging**

Intracellular calcium concentrations ([Ca\(^{2+}\)]_i) were measured by ratiometric imaging techniques (Veterinary Medicine, Iowa State University). Cells were plated onto 22-mm coverslips 36 h before the experiment. Cells were loaded with Fura 2-AM for 60 min at room temperature. The loading solution contained 1 µl of 25% (w/w) Pluronic F-127 and 4 nM of Fura 2-AM diluted in 1 ml of HEPES buffer. The loading solution was removed and the cells were washed twice with HEPES buffer before the coverslips were placed onto a perfusion chamber and connected to a micro pump. The test substances were placed in syringes on a five-valve manifold and applied into the perfusion chamber by the micro pump with a flow rate of 200 µl per min. As a result of the spatial distance between the syringe and the culture in the chamber, there was a time delay between the turning on the valve and onset of the response. Analysis of calcium imaging data was conducted using MetaFluor® software.

**HIV Infections and Cell Viability Assays**

Viral stocks of the infectious molecular clone, pNL4-3, were used for all studies (Adachi *et al.*, 1986). To generate stocks, 15 cm plates of 293T were transfected with pNL4-3 using a calcium phosphate precipitation. Supernatant was collected at 24, 48 and 72 h post transfection. Viral titers were determined as previously described (Reed-Inderbitzin and Maury, 2003) using HeLa 37 cells that express CD4 and the chemokine co-receptors CCR5 and CXCR4 (Platt *et al.*, 1998).

Viral infectivity assays were performed as previously described (Reed-Inderbitzin and Maury, 2003). HeLa37 cells were distributed in a 48-well plate (20,000 cells/well) 24
hours prior infection. The media (DMEM supplemented with 10% FCS and penicillin/streptomycin) was removed and replaced with fresh media containing HIV with *H. gentianoides* extract described above. Between 0.0025 and 0.005 multiplicity of infection (MOI) of virus was added to each well. DMSO concentrations were adjusted so that all wells contained equivalent concentrations and DMSO concentrations were never greater than 1%. All treatments were performed in triplicate. Virus and cells were incubated at 37°C for 40 h.

At the completion of the assay, cells were fixed 10 min with 75% acetone/25% water solution and then immunostained utilizing human anti-HIV antisera (kind gift of Dr. J. Stapleton, Univ. Iowa), followed by HRP-conjugated goat anti-human IgG. 3-amino-9-ethylcarbazole (AEC) was utilized as the HRP substrate. HIV antigen-positive cells were counted and recorded. Wells treated with extracts were compared to the control, infected wells. Data are presented as the number of HIV positive cells in the presence of extract/ the number of HIV positive cells in the absence of extract.

Cell viability studies were performed using ATPLite (Packard Instruments) as previously described (Maury *et al.*, 1994). In brief, the ATPLite kit measures the level of ATP present in the cell population and provides a reliable estimate of cell numbers present. HeLa 37 cells were treated in triplicate with serial dilutions of extracts or equivalent concentrations of the vehicle, DMSO. Cultures were maintained for 40 h and harvested for analysis in parallel with the HIV infection plates. Plates were rinsed once with PBS and lysed per manufacturer’s instructions. Substrate was added and the plates were read in a microtiter plate reader at 430 nm. Data are presented as the ATP present in the well in the presence of extract/the ATP present in the well in the absence of extract.
Microscopic Analysis

Five complete inflorescences of *H. gentianoides* plants (Figure 5 a) were collected and divided into separate vials (two different-aged nodes and internodes per vial) containing 2% paraformaldehyde, 2% glutaraldehyde in 0.1 M cacodylate buffer, pH 7.2. Samples were kept on ice and subjected to 1 h of vacuum (15 psi), then placed at 4°C for up to one week.

The samples were rinsed three times with 0.1 M cacodylate buffer, pH 7.2, before secondary fixation with 1% osmium tetroxide in 0.1 M cacodylate buffer, pH 7.2 for 1 h. Following two rinses with 0.1 M cacodylate buffer, pH 7.2, tissue was en bloc stained with 2% aqueous uranyl acetate for 1 h. Samples were rinsed with deionized water and dehydrated, using the following ethanol solutions for 30 min each: 25%, 50%, 75%, 95%, 100%, and 100%, followed by three changes of pure acetone. The samples were infiltrated with Spurr’s resin using the following ratios of acetone to Spurr’s: 3:1, 1:1, and 1:3. The samples were embedded after two changes of pure Spurr’s and polymerized at 60°C for 48 h.

Samples for SEM (10 flower buds, 10 seed pods, and 10 leaves) were fixed as above up to 100% ethanol. After the final dehydration in 100% ethanol, samples were placed in ultra-pure ethanol. Samples were critical point dried, mounted, sputter coated with palladium gold, and viewed using a JEOL 5800LV SEM. Images were captured digitally using the SIS ADDA II system.

For LM and TEM thick (1 μm) and thin (60 nm) sections were cut from at least 20 flower buds and leaves using a Diatome diamond knife on a Leica UC6 ultramicrotome and stained with 5% uranyl acetate in water for 30 min and in lead citrate for 25-30 min (Sato, 1968). Approximately 35 glands were observed using TEM in total.
Leaves were cleared (removes pigments and soluble cellular contents) by placing in three changes of 90% ethanol for 25 min each at 90°C. The leaves were rehydrated in 70% ethanol at room temperature. The rehydrated samples were placed in a modified Hoyer’s clearing solution (30 g gum arabic, 20 ml glycerol, 150 g chloral hydrate, 60 ml distilled water) (Ilarslan et al., 1997; Liu and Meinke, 1998) for 2 - 4 d.

Anthers were cleared using a technique was modified from Zindler-Frank (1974) to retain the crystals. Anthers and buds were dissected out of floral buds and placed into 95% ethyl alcohol (EtOH) overnight to remove most chlorophyll. Hydrated to deionized water via a graded ethanol series, then placed into 2.5% sodium hypochlorite for 8 h. Cleared samples were rinsed in deionized water (15 min), then dehydrated in an ethanol series, about 15 min each step: 50%, 70%, 95%, 100%, 100% , and further cleared using benzyl alcohol/benzyl benzoate (1:2 v/v, for 2-3 d). After placing the tissue in pure benzyl benzoate, samples were mounted on glass slides in a pure solution of benzyl benzoate, covered with a cover slip, and sealed using clear nail polish. All prepared slides were viewed between crossed polarizers to enhance crystal visibility, and selected images were captured digitally using a Zeiss digital camera.

Crystals were isolated from both fresh (for LM) and critical point dried plant specimens (for SEM). The druses could be easily separated during the maceration of tissue sections in water and crystals were mechanically freed with the help of dissection needles and surgical blades.

Light microscopy images were captured on a Zeiss compound microscope using bright field or polarized light (light is only captured by the analyzer if it has been altered by an object, such as crystal) with a Zeiss Axiocam digital camera. TEM images were captured
on a JEOL 2100, 200 kV scanning transmission electron microscope with a GATAN high resolution 2k x 2k digital camera. Stereomicroscope images were captured using an Olympus stereo microscope with an attached Zeiss Axiocam digital camera.

Acknowledgements

We thank Mark Widrlechner and Joe-Ann McCoy from the USDA-ARS, North Central Regional Plant Introduction Station (112 Plant Introduction Farm, Iowa State University, Ames, IA 50011) for providing seeds of characterized Hypericum accessions. We are grateful to Ann Perera and Basil J. Nikolau at the W. M. Keck Metabolomics Facility, at Iowa State University; and Harry Horner, Tracey Pepper, and Randy DenAdel at the Microscopy and NanoImaging Facility for providing the analytical instrumentation and for valuable suggestions. This research was made possible by grant P01 ES012020 from the National Institute of Environmental Health Sciences (NIEHS) and the Office of Dietary Supplements (ODS), National Institutes of Health (NIH), grant 9P50AT004155-06 from the National Center for Complementary and Alternative Medicine (NCCAM) and ODS, NIH, and grant EEC-0813570 National Science Foundation Engineering Research Center for Biorenewable Chemicals. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NSF, NIEHS, NCCAM, or NIH.

References


Figure 1
Figure 1. *Hypericum gentianoides* plants. Plants were established from seeds harvested from just-matured pods, and immediately sown in soil (65% peat moss, 30% perlite, and 5% mineral soil) and grown in a greenhouse at 22 – 25°C under natural light. (a) Representative flowering *H. gentianoides* plant used in metabolite extractions and morphological studies, having vegetative growth (v) and just starting to flower (f). (b) Mature flowering plant with tall, spindly growth habit characteristic of *H. gentianoides* once flowering starts and vegetative growth ceases. (c & d) Cleared inflorescence leaves, viewed between crossed polarizers, which removes light from the system unless it is altered by the specimen, and showing bright white venation, with glands located in the open areas between veins (white arrow). Bars = 50 μm. (e) Bar chart showing mean gland occurrence on various organs of *H. gentianoides*. Vegetative leaves and sepals contain significantly more glands than leaves on inflorescences or stems (* denotes a t-test p-value < 0.01). The mean number of glands is not significantly different comparing vegetative leaves and sepals. Results shown are the mean ± SE for the number of glands found on 10 individual organs. For internodes: one side was counted and multiplied by four to estimate total number of glands.
Figure 2. Diversity in flower size from six *Hypericum* species. *Hypericum* species generally have yellow flowers (Ernst, 2003) and as shown, there is considerable diversity in flower size. *Hypericum gentianoides* has the smallest flower of the six species shown and *H. perforatum* is the most widely studied species of *Hypericum*. 
Figure 3. *Hypericum gentianoides* extract causes a transient cytosolic Ca\(^{2+}\) increase in human embryonic kidney cells and primary cultures of rat hippocampal astrocytes. (a) Time course of calcium change in HEK293 cells in response to repeated application of *H. gentianoides* extract (50 µg dry weight extract/ml DMSO). Fresh shoots were extracted with 100% methanol (Hillwig *et al.*, 2008), dried, and re-dissolved in 100% DMSO before application. Data presented are traces from 25 cells in a single biological experiment, representative of three independent experiments. Each line corresponds to a trace from one HEK293 cell. Cells treated with control (100% DMSO) did not evoke a response (data not shown). (b & c) Pseudocolor images of calcium transients following treatment with DMSO (b) and treatment with *H. gentianoides* extracts (c). (d) Table quantifying calcium changes measured for HEK 293 cells and rat astrocytes treated with *H. gentianoides* extracts.
Figure 4

(a) Graph showing the effect of extract concentration on cell viability and HIV infection. The x-axis represents extract concentration (µg/ml) ranging from 1 to 100, and the y-axis represents % Control (% of control values). Two lines are plotted: one for cell viability and one for HIV infection.

(b) Bar graph showing the % control of cell viability and HIV infection at different extract concentrations for three different strains: Ames 28015 - Dry, Ames 27480 - Fresh, and Ames 28015 - Dry. The x-axis represents extract concentration (µg/ml) and the y-axis represents % control.
Figure 4. Methanolic extracts from *Hypericum gentianoides* reduce HIV infection in HeLa37 cells. Cultured HeLa37 cells assays for viral infectivity (Reed-Inderbitzin and Maury, 2003) with three biological replicates were used to test the ability of *H. gentianoides* extracts to inhibit HIV infectivity of HeLa37 cells. (a) Varying concentrations of an *H. gentianoides* methanolic extract was applied along with HIV to HeLa37 cells to test the ability of the extracts to inhibit HIV infection. Cell viability was measured using the ATPLite kit. (b) Multiple concentrations of *H. gentianoides* methanolic extracts prepared from plants of two different accessions (Ames 28015 and Ames 27480) obtained from the USDA-ARS, North Central Regional Plant Introduction Station were tested for the ability to inhibit HIV infection in HeLa37 cells. Significance was determined using a t-test to make pair-wise comparisons for each organ type (* p-value < 0.01, # p-value < 0.05).
Figure 5. Relative concentrations of acyl-phloroglucinols in different sections of the *Hypericum gentianoides* inflorescences. (a) Representative inflorescence showing tissue collected from *Hypericum gentianoides* for metabolite and microscopic analyses. This image shows the locations at which inflorescences were dissected and collected for metabolite extraction. (b) Values are normalized to tissue weight. Results shown are the mean ± SE of two biological replicates each containing pooled material from 10 inflorescences. The highest concentrations of saroaspidin A, uliginosin A, and hyperbrasilol C are found in flower buds and internode 2. Four additional putative diacylphloroglucinols (as determined by LC-MS-MS data, (Hillwig, 2008)) are also shown, indicated by their retention time. Significance was determined using a t-test to make pair-wise comparisons for each organ type (* p-value < 0.01, # p-value < 0.05).
Figure 6. ESI-MS [M-1] infusion spectra from contents *H. gentianoides* glands show they contain almost exclusively saroaspidin A (445 m/z), uliginosin A (500 m/z), and hyperbrasilol C (553 m/z). Pulled capillary needles were used to puncture 250 individual glands on *H. gentianoides* leaves and retrieve the contents. The pooled glandular material was dissolved in 500 µl of methanol and directly infused into an Agilent Ion Trap 1100 electrospray-ionization mass spectrometer.
Figure 7. *Hypericum gentianoides* floral bud organs. (a) Longitudinal section of a young flower bud containing young stigma, anthers and adjacent developing bud. The outer protective leaves contain a large number of glands. (b) Longitudinal section of mature flower bud showing sepals, petals, and anthers. Large glands are present in the sepals (Figures 8 – 11 detail the ontology of these glands). (c – d) Cross sections through anthers at the microsporogenesis stage, showing connective tissue containing a single large gland partially surrounded by calcium oxalate crystal druses. White and black arrows denote the location of calcium oxalate crystals, while the red arrow shows gland location. (c) Light microscopic images showing the anther connective tissue with one gland (red arrow) and calcium oxalate crystal druses (black arrows). (d) The same section as in (c), with calcium oxalate crystal druses visualized between crossed polarizers (Horner and Wagner, 1992). (e) Cleared whole anthers observed using LM. White arrow points to crystals. (f, h - i) Crystals were isolated from macerated tissue using a surgical blade and dissecting needle for LM and SEM analysis. (f) Isolated crystal druses (large aggregates of multiple crystal facets) observed under crossed polarizers. White arrow points to a single crystal. (g) Crystal contained within a cell observed under SEM. (h - i) Isolated crystals observed using SEM. (j – l) SEM images of *H. gentianoides* anther. (j) Red arrow points to a gland containing secretory product. Calcium oxalate crystals appear in the surrounding connective tissue (small white arrows). (j) Intact anther with red circle denoting to gland location. (k) Anther showing pollen in locules, a gland (red arrow) and crystals (white arrows).

Bars: (a - b) 100 μm, (c -d) 20 μm, (e) 50 μm, (f – g) 5 μm, (h) 10 μm, (i) 5 μm, (j – l) 25 μm

Figure Abbreviations. an = anther, b = bud, g = gland, L = outer protective leaves, p = petal, s = sepal
Figure 8. Development of *Hypericum gentianoides* glands. Thick sections (1 μm) of fixed and embedded sepal tissue, stained with 1% toludine blue O examined by bright field LM. (a) Stage 1 gland. White arrow points to a young gland filled with cytoplasmically-dense secretory cells. The cells surrounding the gland contain osmium-stained material. (b) Stage 2 gland. Secretory cell cytoplasm is still dense, but the lumen is starting to form. (c) Stage 3 gland. A droplet of lightly osmium-stained material within the lumen. Osmium-stained material is still present in cells surrounding the gland. (d) Stage 4 gland with a central lumen and slightly less osmium-stained material is present in surrounding cells. (e) Stage 5 gland secretory cells contain more vesicle-like structures and surrounding cells stain less intensely with osmium. (f) Stage 6 gland. Peripheral cells appear empty and secretory cells are a mere line.

Bars: 20 μm.

Figure Abbreviations. l = lumen, pc = peripheral cells, sc = secretory cells
Figure 9. Young (stages 1 & 2) Hypericum gentianoides glands. (a) Young gland with cytoplasmically-dense secretory cells and large peripheral cells. Large vacuoles containing dark osmium-stained material are prominent in peripheral cells (b) Vesicle-like structures form in the cytoplasm of secretory cells along cell walls. (c) The cell walls of secretory cells appear degraded in stage 2 glands (d) Stage 2 gland secretory cell. Vesicle-like structures within secretory cells contain material stained with osmium. The dense cytoplasm of secretory cells contains plastids with tubuli, ghost-like mitochondria, and a network of ER and ribosomes. (e) Tubuli containing plastids and vesicles of secretory cells. Osmium-bodies and vesicle-like structures are present in the cytoplasm of secretory cells. (f) Peripheral cells contain chloroplasts with grana and starch. The vacuole of the peripheral cell is packed with material densely stained with osmium. The cell walls separating the peripheral cells and secretory cells are very thick, while walls separating the secretory cells from the lumen are thinner. Material stained with osmium is present in vesicle-like structures.

Bars: (a) 10 μm, (b – f) 2 μm

Figure Abbreviations. c = chloroplast, cw = cell wall, m = mitochondria, n = nucleus, no = nucleolus, ob = osmium-body, p = plastid, pc = peripheral cells, sc = secretory cells, vs = vesicle-like structures.
Figure 10. Developing (stages 3 & 4) *Hypericum gentianoides* glands. (a) Developing glands have an enlarged lumen filled with secretory material. The secretory cells are pushed up against the peripheral cells. The cytoplasm of both the secretory cells and peripheral cells is becoming less dense. (b) Stage 3 gland. Cytoplasm of secretory cells contain more vesicle-like structures compared to secretory cells found in young glands. Vesicle-like structures lightly stained with osmium accumulate against a thinned cell wall. (c) Stage 3 gland. Large osmium-bodies are present in secretory cells. Fibrillar material can be seen coming off the cell wall on the luminal side of the secretory cell. (d) Stage 3 gland. Secretory cell filled with vesicles-like structures. Vesicle-like structures also appear along the thinned secretory cell wall. (e) Cytoplasmically-dense secretory cells of stage 3 glands still contain specialized plastids. The peripheral cells contain darkly-stained material and chloroplasts. (f) Stage 3 secretory cell containing many plastids and osmium-bodies. (g) Stage 4 gland. Developing secretory gland with a less-dense cytoplasm and several vesicle-like structures containing particulate matter. (h) Vesicle-like structures within a secretory cell, having a reduced cytoplasm, contain osmium-stained material. (i) Stage 4 gland. Enlarged vesicle-like structures containing osmium within secretory cells. (j) Stage 4 gland. Secretory cells with reduced cytoplasm and peripheral cells with enlarged vacuoles containing osmium-stained material. Bar = 10 μm. (k) Stage 4 gland. Secretory cell cytoplasm is degenerated and osmium-stained vesicle-like structures are found along a thinned cell wall. The cytoplasm of the peripheral cell is degraded and the chloroplasts are degenerated. Note that the cell wall separating the peripheral cell from the secretory cell is much thicker and more organized than the wall separating the secretory cell from the lumen. (l) Stage 4 gland. Secretory cell cytoplasm is filled with plastids containing tubuli and is peppered with ribosomes.

Bars: (a) 20 μm, (b – i) 2 μm, (j) 10 μm, (k – l) 2 μm

Figure Abbreviations. c = chloroplast, cw = cell wall, f = fibrillar material, m = mitochondria, ob = osmium-body, p = plastid, pc = peripheral cells, sc = secretory cells, v = vacuole, vs = vesicle-like structures.
Figure 11. Mature (stages 5 & 6) *Hypericum gentianoides* glands. (a) Stage 5 gland. Portion of a secretory cell with an elongated vacuole filled with osmium-stained material. Chloroplasts can be found in peripheral cells. (b) Stage 5 gland. Cytoplasm of secretory cell found in a mature gland. Plastids containing tubuli are still present, but cytoplasm is less dense than cytoplasm observed in secretory cells from younger glands. (c) Stage 5 gland. Cell wall separating peripheral and secretory cell walls are much thicker than walls in between secretory cells. (d) Stage 5 gland. Vesicles-like structures with osmium-stained material within a secretory cell. (e) Stage 5 gland. Large vesicle-like structure located in a cytoplasm-containing secretory cell, while a neighboring peripheral cell contains little to no cytoplasm. (f) Stage 5 gland. Secretory cell with cytoplasm and plastids containing tubuli. Material can be seen in the lumen. (g) Mature glands contain a large central lumen and secretory cells are appressed to peripheral cells. Secretory cells are practically devoid of cytoplasm and contain a large vacuole that contains droplets of osmium-stained material. (h) Stage 6 gland. Droplets of osmium-stained product are present in the large vacuoles of secretory cells. (i) Stage 6 gland. Osmium-stained droplets in secretory cell vacuole. Note fibrillar material on the lumen side of secretory cell wall and secretory product in lumen. (j) Stage 6 gland. Secretory cell vacuole contains osmium-stained droplets. Fibrillar material is present along the secretory cell wall on the lumen side. (k) Stage 6 gland. Secretory cells and peripheral cells contain large vacuoles in mature glands. In mature glands, more densely stained material is localized to secretory cells.

Bars: (a – f) 2 \( \mu m \), (g) 20 \( \mu m \), (h) 10 \( \mu m \), (i – j) 2 \( \mu m \), (k) 10 \( \mu m \), (l) 2 \( \mu m \)

Figure Abbreviations. c = chloroplast, cw = cell wall, f = fibrillar material, m = mitochondria, p = plastid, pc = peripheral cells, sc = secretory cells, v = vacuole, vs = vesicle-like structures.
CHAPTER 4. GENERAL CONCLUSIONS

This dissertation has shown that, surprisingly, antisense-ACLA plants rescued by malonate do not exhibit large transcriptional or metabolic changes, indicating the phenotypic reversion observed is primarily under post-transcriptional control. This dissertation has also shown that translucent glands of *H. gentianoides* develop schizogenously and they contain saroaspidin A, uliginosin A, and hyperbrasilol C. Methanolic extracts containing saroaspidin A, uliginosin A, and hyperbrasilol C have the ability to increase intracellular calcium levels in HEK293 cells and cultured primary rat astrocytes. These same extracts have the ability to inhibit the infectivity of HIV in viral infectivity assays.

*Antisense-ACLA plants rescued by malonic acid do not have global transcriptomic or metabolomic changes*

A major phenotypic reversion is caused by the application of malonic acid to antisense-ACLA plants (Fatland et al., 2005). Malonate application is able to return starch and anthocyanin to WT-like levels, as well as convert the bonsai-like growth habit and cuticular wax composition of antisense-ACLA plants back to WT-like levels. It would seem that a large number of changes must occur at the transcriptional and metabolomic levels to facilitate the observed phenotypic reversion. Surprisingly, the number of genes with altered expression that differ between antisense-ACLA water- and malonic acid-treated plants is quite small, while large numbers of genes accumulate differentially when comparing antisense-ACLA and WT plants.
Genotype specific gene expression patterns

Starch and anthocyanin biosynthesis in antisense-ACLA plants is under post transcriptional control. When antisense-ACLA plants are given malonic acid, the starch and anthocyanin hyper-accumulation phenotype disappears and levels revert back to WT-like. It would be expected that RNA levels for genes related to starch and anthocyanin biosynthesis would also revert to WT levels. This is not the case, as the expression of anthocyanin and starch biosynthesis genes is higher in antisense-ACLA than WT plants regardless of treatment, water or malonic acid.

In total, seven transcripts from genes known to be involved with anthocyanin biosynthesis are up-regulated in antisense-ACLA plants, regardless of treatment. Anthocyanin biosynthetic genes up-regulated include chalcone synthase (CHS, AT5G13930), the first committed step of anthocyanin biosynthesis and flavanone 3-hydroxylase (F3H, AT3G51240), which has been shown to physically interact with CHS (Owens et al., 2008). TRANSPARENT TESTA 19 (TT19, AT5G17220), a gene whose product is responsible for depositing anthocyanidin 3-glucosides into the vacuole (Kitamura et al., 2004), is also up-regulated in antisense-ACLA plants. Presumably, the increased expression of CHS leads to the production of more enzyme which in turn leads to increased anthocyanin accumulation in antisense-ACLA plants. It is interesting that the expression level of CHS does not return to normal levels in antisense-ACLA plants treated with malonic acid, indicating a signal other than lack of malonyl-CoA is responsible for the regulation of its expression.

Up-regulated starch synthesis genes include ADP-glucose pyrophosphorylase 3 (APL3, AT4G39210) and 4 (APL4 AT2G21590), a putative starch synthase (AT1G32900).
and a putative sucrose synthase (AT3G43190). APL3 and APL4 are regulatory subunits of ADP-glucose pyrophosphorylase that are induced by sucrose and trehalose. Sucrose and trehalose induction of APL3 and APL4 decouple starch synthesis from allosteric control by 3-phosphoglycerate, causing an increase in starch synthesis (Crevillen et al., 2005). While application of malonate causes starch to accumulate at WT-like levels, the expression of starch synthesis related transcripts is unchanged. While malonate is sufficient to return starch synthesis to normal levels, it does not influence the expression of APL3 or APL4. In the metabolome several unidentified sugars accumulate to lower levels in antisense-ACLA plants.

Antisense-ACLA plants have altered cuticular wax structure and content (Fatland et al., 2005). Levels of C29 alkanes and secondary alcohols are reduced in antisense-ACLA plants. Expression of CER1 (AT1G02205), an octadecanal decarbonylase suspected of producing odd-numbered carbon length alkanes and secondary alcohols, is presumably increased in antisense-ACLA plants to make up for the deficiency of C29 alkanes and secondary alcohols.

The long chain fatty acid docosanoic acid accumulates to higher levels in antisense-ACLA plants treated with water than malonic acid. It is surprising that a long chain fatty acid (elongated in the cytosol using acetyl-CoA derived malonyl-CoA) would accumulate to high levels in antisense-ACLA plants. DAISY (AT1G04220) encodes an active docosanoic acid synthase and it accumulates to higher levels in antisense-ACLA plants compared to WT plants. The expression of DAISY is not altered in antisense-ACLA plants treated with malonate even though docosanoic acid levels are decreased.
In antisense-ACLA plants the expression of ACLA-2 and ACLA-3 is decreased compared to WT regardless of treatment. As shown in Chapter 2, Figure 2 A, genes for ACL subunits are co-expressed, so it follows that ACLA-2 and ACLA-3 transcripts are down-regulated in antisense-ACLA plants.

**Gene expression in response to malonate treatment**

The expression of several genes is influenced by malonate treatment. Twenty-eight transcripts accumulate with a pattern of SUD. These transcripts include the thalianol operon genes (AT5G47980, BADH acyltransferase; AT5G47990, THAD; AT5G4800, THAH; AT5G48010, THAS). Thalianol usually accumulates in roots, but in Arabidopsis with leaf accumulation of thalianol, plants are smaller than normal (Field et al., 2008). At this time it is unknown if antisense-ACLA plants also accumulate thalianol in leaf tissue.

The expression of two floral identity genes, SEPALLATA 2 (SEP2, AT3G02310) and APPETALA 3 (AP3, AT3G54340) in Arabidopsis rosette leaf tissue is surprising. The expression of a SEP2 homologue has been reported in Gerbera leaf tissue (Malcomber et al., 2005). Antisense-ACLA plants do have a longer life span, are more branched, and have shorter petals than WT. These differences could be influenced by the accumulation of SEP2 and AP3 in leaf tissue.

The ability of malonate to rescue antisense-ACLA plants indicates the presence of a malonyl-CoA synthetase in Arabidopsis. Of the transcripts having altered expression in antisense-ACLA plants treated with malonic acid, none looked like candidates for the malonyl-CoA synthetase.
Hypericum gentianoides translucent glands contain the bioactive compounds saraoaspidin A, uliginosin A, and hyperbrasilol C

Gland development

Hypericum gentianoides plants have translucent glands, containing the acyl-phloroglucinols, saraoaspidin A, uliginosin A, and hyperbrasilol C, on leaves, stems, and sepals. The development of these translucent glands is schizogenous. Very young H. gentianoides glands at stages 1 and 2, are characterized with peripheral cells containing darkly stained material; and cytoplasmically dense secretory cells (CHAPTER 3; Figure 8 A-B). Vesicles can be found along secretory cell walls prior to cell wall breakdown, presumably carrying pectinases and/or cellulases (CHAPTER 3; Figure 11 C). The secretory cell cytoplasm of stage 1 and 2 glands is filled with ER, ribosomes, specialized plastids with multidirectional tubuli, and degenerating mitochondria (CHAPTER 3; Figure 11).

As glands age (stages 3 and 4), fibrillar material is observed along secretory cell walls on the luminal side (CHAPTER 3; Figure 11). This material could be cell wall debris or secretory product being released into the lumen via diffusion. In stage 3 and 4 glands, osmium-stained vesicles are present along secretory cell walls, presumably depositing secretory product into the lumen (granulocrine secretion) (CHAPTER 3; Figure 11). Mature glands (stages 5 and 6) have much reduced secretory cells. These cells are appressed to along the peripheral cells by the pressure of luminal contents.

Cells surrounding glands stain darkly with osmium in stages 1 and 2 (CHAPTER 3, Figure 8 A-B), possibly containing phloroglucinols precursors for secretory material synthesis. Over time (stages 3 through 6) osmium staining of surrounding cells decreases
At the ultrastructural level, osmium stained material is located in peripheral cell vacuoles in young stages 1 and 2 glands (CHAPTER 3; Figure 10) and this material lightens in mature glands, stages 5 and 6 (CHAPTER 3; Figure 12). In mature glands (stages 5 and 6), osmium-stained material is located in enlarged vacuoles (CHAPTER 3; Figure 12), and this material is likely a collection of the darkly stained material found in vesicles of younger secretory cells (CHAPTER 3; Figures 10, 11). The change in the intensity of osmium staining indicates that the chemical composition of a compound has changed. In the case of the peripheral cells, it is likely the reduced staining indicates a reduction in the amount of material stored in the cells. Direct evidence is needed to show that this is indeed what is happening (or “what happens”) during secretion into the translucent glands of *H. gentianoides*. An approach such as GALDI-MS (Cha et al., 2008) could help determine where saroaspidin A, uliginosin A, and hyperbrasilol C are actually synthesized. Ceratin features of secretory cells within the gland indicate high levels of biosynthesis, i.e., ER network, specialized plastids, and cytoplasm peppered with ribosomes (Fahn, 1979).

**Bioactivity**

Methanolic extracts from *H. gentianoides* have the ability to cause an increase in intracellular calcium levels in HEK293 cells and primary rat astrocytes. These are two very different cell types, indicating that compound(s) from the *H. gentianoides* extracts can alter cellular calcium homeostasis in a variety of ways. *Hypericum gentianoides* methanolic extracts also have the ability to reduce the infectivity of HIV in assays, with stronger cytotoxicity coming from fresh plant material. These results, combined with the previously studied ability of *H. gentianoides* methanolic extracts to reduce LPS-induced inflammation in
RAW 264.7 mouse macrophage cells (Hillwig et al., 2008), show importance of further study of these bioactivities.

References


APPENDIX A. EXPLORASE: MULTIVARIATE EXPLORATORY ANALYSIS AND VISUALIZATION FOR SYSTEMS BIOLOGY

A paper published in the Journal of Statistical Software

Michael Lawrence¹, Dianne Cook², Eun-Kyung Lee³, Heather Babka², Eve Syrkin Wurtele²

Abstract

The datasets being produced by high-throughput biological experiments, such as microarrays, have forced biologists to turn to sophisticated statistical analysis and visualization tools in order to understand their data. We address the particular need for an open-source exploratory data analysis tool that applies numerical methods in coordination with interactive graphics to the analysis of experimental data. The software package, known as explorase, provides a graphical user interface (GUI) on top of the R platform for statistical computing and the GGobi software for multivariate interactive graphics. The GUI is designed for use by biologists, many of whom are unfamiliar with the R language. It displays metadata about experimental design and biological entities in tables that are sortable and filterable. There are menu shortcuts to the analysis methods implemented in R, including graphical interfaces to linear modeling tools. The GUI is linked to data plots in GGobi through a brush tool that simultaneously colors rows in the entity information table and points in the GGobi plots.

explorase is an R package publicly available from Bioconductor and is a tool in the MetNet platform for the analysis of systems biology data.

Keywords: bioconductor, bioinformatics, microarray, graphical user interface, exploratory data analysis, interactive graphics, visualization, metabolomics, proteomics.

¹ Fred Hutchinson Cancer Research Center
² Iowa State University
³ Ulsan University
1. Introduction

In recognition of the need for biologists to analyze multidimensional, high-throughput data, we have developed a software tool with the following goals:

• Support exploratory analysis of experimental data through the integration of numerical methods and interactive graphics.
• Leverage biological information in the analysis of experimental data.
• Provide a graphical user interface (GUI) on top of statistical methods so that they are accessible to biologists.

High-throughput experiments have become commonplace in biology. The popular microarray measures the levels of tens of thousands of gene transcripts at once. GC-MS and other analytical methods currently have the potential to detect hundreds of metabolite levels, providing a snapshot of the metabolism in a cell. The explorase package has been developed for analyzing datasets with measurements on tens of thousands of biological entities, such as genes or metabolites. The explorase interface and analysis algorithms are designed for experiments with up to approximately 50 samples (columns in the experimental data matrix).

The untargeted nature of high-throughput experiments pairs well with an approach to data analysis that remains open to the unexpected and allows the analyst to form hypotheses during analysis. The intent is to facilitate the search for interesting features in the data, such as differentially expressed genes or metabolites that follow a similar pattern. This approach is known generally as exploratory data analysis and is the main philosophy behind the design of our software tool.
Interactive graphics are generally useful in exploratory analysis, and they are particularly applicable to analyzing experimental data for several reasons:

- Experimental datasets are multivariate and so benefit from the ability to view different combinations of variables in different ways, simultaneously.
- Results from numerical methods can be interpreted and validated by relating them back to the original data through visual cues.
- Linked interaction can relate individual measurements to the biological system by integrating plots of experimental data with drawings of biochemical networks.

There is a number of software projects for analyzing data from high-throughput biological experiments. One of these is Bioconductor (Gentleman et al. 2004), a free collection of R packages (R Development Core Team 2008) that provide numerical and graphical methods for helping biologists comprehend their data. There are packages in Bioconductor for analyzing and visualizing networks and various types of experimental data, including microarray and mass spectrometry data.

While R and the Bioconductor project provide sufficient support for numerical and general graphical methods, the R graphics system is not designed for interactivity. A specialized application for multivariate interactive graphics, GGobi (Swayne et al. 2003), has been interfaced with R through the package rggobi (Temple Lang and Swayne 2001; Temple Lang et al. 2008). GGobi provides interactive scatterplots, parallel coordinate plots, barcharts and other types of displays. Edges may be displayed in scatterplots in order to draw networks. Interaction modes include linked brushing between plots by identifier and categorical variable, point and edge querying, and pan/zoom.
The complementary nature of R/Bioconductor and GGobi encourages their combined application to the analysis of biological data. However, this is hindered by their lack of accessibility to biologists. All R packages, including those in Bioconductor, are driven through the R language, which facilitates their application to a wide variety of data analysis tasks. The flexibility and expressiveness of a script-driven interface is a double-edged sword, however. Biologists unfamiliar with programming and command-line interfaces struggle to take advantage of R and Bioconductor packages that lack a GUI. GGobi is designed to be flexible and open-ended, with the goal of supporting a wide range of analyses. However, this generality means that the biologist receives no biology-specific guidance during data analysis and visualization tasks.

We have developed a GUI-driven R package named explorase for the graphics-intensive exploratory analysis of biological experimental data. It is a tool in the MetNet platform (http://www.metnetdb.org/, Wurtele et al. 2003, 2007) for systems biology data analysis and is publicly available from Bioconductor (http://www.bioconductor.org/) as of Bioconductor version 2.1.

The explorase package has the following general features:

- A collection of numerical methods, such as distance measures and linear models, implemented in the R language (R Development Core Team 2008).
- Linked data plots based on the GGobi tool for multivariate interactive graphics (Swayne et al. 2003).
- A GUI, designed in collaboration with biologists, that integrates numerical methods with graphics and provides general features such as the loading and subsetting of data.
Section 2 presents an overview of the software and the next section describes the numerical and graphical methods it provides. The layout of its GUI is detailed in Section 4. This is followed by a tutorial that guides the user through a hypothetical analysis. Finally, the paper will conclude by discussing the future of explorase.

2. Overview

As its name suggests, explorase is designed to facilitate the exploratory analysis of biological data by combining the numerical methods of R (and Bioconductor) with the graphical methods of GGobi. Most users will only interact with explorase through its GUI, and the GUI and plots of GGobi. The explorase GUI is shown in Figure 1.

The user begins an explorase session by loading information describing one or more experiments. This includes the actual experimental measurements, as well as metadata and other supporting information. Each type of information is described below:

**Experimental data matrix:** Measurements of the levels of transcripts, metabolites or some other biological entity for every sample in the experiment.

**Experimental design matrix:** Description of experimental design, factors such as genotype or time. Each sample is labeled by the factor levels of the experimental design.

**Entity annotation matrix:** Annotations, such as GO terms, of biological entities measured in the experiment.

**Entity list:** User-defined lists of entities of interest, such as lists of metabolites in a common pathway or co-expressed genes.

While no particular type of information is required to run explorase, each feature has specific data requirements. For example, the linear modeling features require an experimental design matrix and the experimental data.
The data from a set of related experiments are organized into a *project*. The user creates a project by importing individual data files into an empty explorase session. In future sessions, the user only needs to load the project to begin the analysis. Only one project may be loaded per session, and a project may contain at most one experiment for each type of biological entity. The entity types built into explorase are genes, metabolites and proteins; expert users may define custom types through the R command line interface. The distinction between entity types is currently only for organizational purposes, though it may gain meaning in the future.

The files in a project are physically organized into a directory (folder) in the file system. The user need only specify the directory to load every file (i.e., experimental data files and metadata files), in the project at once. The format of every file is CSV, which is compatible with most spreadsheet applications. The type of information contained in a file and the type of biological entity the information describes are indicated by the filename extension.

Once a dataset and related information is loaded into explorase, the user may proceed with the analysis by performing operations such as:

- Subset the observations by criteria such as the minimum fold change and maximum variance across replicates.
- Check the quality of the data with GGobi graphics, such as scatterplots, scatterplot matrices and parallel coordinate plots.
- Browse entity metadata and analysis results in a filterable, searchable and sortable table.
• Color rows in the entity information table and the corresponding points in GGobi plots using the brush tool.

• Detect differentially expressed genes through the graphical interface to the limma package (Smyth 2005) and view their profiles in a GGobi parallel coordinate plot.

• Find patterns in the data through hierarchical clustering and the explicit pattern query tool.

• Export analysis results and lists of interesting entities as CSV files.

3. Methods

3.1. Numerical methods

The numerical methods in explorase are designed to assist in finding biological entities with patterns that depend on experimental conditions or are similar to the pattern of a given entity or user-specified pattern. It is also possible to cluster entities according to a selected distance measure.

Finding entities with interesting patterns

One means of finding interesting patterns with explorase is to compare two replicates or replicate means by a selected distance measure. The supported distance measures are difference, residuals from regressing one condition against the other, angle between the diagonal and the line from the origin to the point in the scatterplot of the two variables, and the Mahalanobis distance (e.g., Johnson and Wichern 2002). Of these, the difference is the simplest and the most often applied to transcriptomics data. A distance measure may be used to check the agreement between two replicates or to evaluate differences in entity levels between treatments using replicate means.
Linear modeling is a more sophisticated technique for estimating the effects of experimental conditions on each entity. The explorase GUI includes an interface to limma (Smyth 2005) that fits a linear model to each entity and shrinks the variance using empirical Bayes analysis to yield p value estimates that tend to correspond to what a biologist considers interesting: entities with relatively high basal levels as well as significant difference across conditions. The significance of each user-selected experimental design factor, as well as their interactions, is estimated without the user needing to manually specify any contrasts. The output of the limma tool may consist of the raw p values, p values corrected by FDR or another method, the corresponding F statistics, the contrast coefficients and the fitted values. The limma interface has an advanced feature for fitting contrasts that evaluate whether an entity pattern is linearly or quadratically dependent on time in time-course experiments.

There is also a separate polynomial model for estimating the effect of time. Its critical difference from limma is that it accounts for the order of the time points and thus is likely more realistic for time-course experiments. The user also can apply this polynomial model to evaluate interaction effects between time and other factors. The output contains the p value and coefficient for each effect, as well as the F statistic and the sum of squares error for the overall model.

*Searching for entities with specific patterns*

The explorase package offers several distance measures for comparing entity patterns. These include cosine angle (uncentered correlation), Euclidean, Pearson correlation and Canberra (e.g., Johnson and Wichern 2002). The Euclidean distance is usually not of interest, as it is based on the magnitude of the levels, not their pattern; however, it may be useful for finding entities that are present at similar levels over the course of the experiment.
The Pearson correlation disregards magnitude, so it may be useful for identifying entities with similar patterns regardless of their levels. This could, for example, help identify gene regulatory interactions, where the increase (or decrease) in one transcript results in the decrease (or increase) in another. In contrast, the cosine angle distance considers both pattern and magnitude and thus may be useful if the biologist wishes to focus on entities present at similar levels while still considering the pattern. If the level of the query entity is relatively high (i.e., above background noise), the cosine angle measure reduces the number of hits against entities that are present only at low (i.e., background) levels and may thus be considered uninteresting. The Canberra distance may be particularly useful for metabolomics data as it is numerically stable when faced with zero values, which result from the common practice of imputing non-detects as zeros.

The user may also search for an explicit pattern by specifying “Up”, “Down” or “Same” for each transition between adjacent samples in a user-specified list. The entity patterns are matched to the query by denoting a transition as “Up” if the transition value (difference between the pair of samples) is above the $q$ quantile and “Down” if the value is below the $1 - q$ quantile. Everything between the quantiles is marked as “Same”. The parameter $q \in [0, 1]$ is specified by the user with a slider.

**Clustering of entities**

For clustering the observations, explorase performs agglomerative hierarchical clustering using Ward’s linkage method (e.g., Johnson and Wichern 2002). The user can choose from the several distance measures: cosine angle, Euclidean, Pearson correlation and Canberra. These are the same distance measures described for comparing entity patterns
against a query entity pattern. A very similar question is asked when clustering, so the same
criteria apply for selecting a distance measure. One of the major differences between
comparing entity patterns and clustering is that in clustering the distance is calculated
between every pair of entities rather than between one entity and the others. Given the size of
high-throughput datasets, it is computationally intensive to calculate the pairwise distance for
every entity in the experiment, and the results can be difficult to interpret. Thus, it is
recommended that the user select a small subset of entities for clustering. explorase supports
efficient clustering of up to approximately 50 entities.

3.2. Graphical methods

The graphics in explorase are interactive and designed for exploratory data analysis.
The data plots are displayed by GGobi. For visualizing the experimental data, explorase
provides scatterplots, scatterplot matrices, parallel coordinate plots and histograms.

Scatterplots

The scatterplot, or the basic X vs. Y plot, is the most generally applicable explorase
plot type. For example, the user may compare two samples, such as a pair of replicates, or
two conditions, averaged over the replicates. In this way, the scatterplot is the graphical
equivalent of distance measures for comparing the levels of an entity between two samples or
conditions. The scatterplot is also useful for interpreting analysis results. For example,
comparing the p value for a limma effect with its coefficient, as in Figure 2, is particularly
effective for detecting patterns that are both significant and of high amplitude.

Scatterplot matrices

The scatterplot matrix is a symmetric grid of scatterplots. Every row has the same
variable on the vertical axis and every column shares the variable on the horizontal axis.
Along the diagonal, the X and Y variables are the same, so histogram is displayed for that variable. Scatterplot matrices are useful for obtaining an overview of the data, because multiple variables are compared at once. One common application is the comparison of replicate sets during data quality checking.

**Parallel coordinate plots**

The parallel coordinate plot displays a profile of each entity across a sequence of variables, normally the samples or replicate averages. Due to the large number of entities in microarray, proteomics and metabolomics experiments, the profiles tend to be overplotted to the extent that individual profiles are not recognizable (e.g., the grey profiles in the parallel coordinate plot in Figure 2). Linked brushing helps to overcome this problem. For example, as shown in Figure 2, the user could brush outlying points in the scatterplot of p value versus coefficient to highlight the corresponding profiles in the parallel coordinate plot. This permits the visual validation of patterns deemed significant by numerical methods.

**Histograms**

The interactive histogram is another tool for interpreting analysis results and checking data quality. The user can view profiles that are outliers with respect to a statistic by brushing outlying points in the histogram of the statistic. For example, assume the user has brushed a particular gene of interest and wishes to view profiles that are similar to that of the chosen gene. The user could calculate a distance measure between the gene and the others and compare profiles in a parallel coordinate plot by brushing the outlying points (with a new color) in a histogram of the distances.

### 4. GUI features
The main GUI of explorase, shown in Figure 1, consists of three panels (Entity information, Samples/Treatments and Lists/Pathways), a toolbar and a menubar. The primary design considerations for the GUI are simplicity and usability. There is no attempt to completely map the features of the underlying tools to the explorase GUI. Rather, the GUI supports a subset of the features most useful in the analysis of high-throughput transcriptomics, proteomics and metabolomics data and augments this subset with shortcuts and conveniences for biological data analysis. The GUI has been designed in collaboration with biologists, to help ensure that explorase is accessible to those who are performing the experiments and generating the data.

4.1. Main panels

**Entity information panel**

The largest panel contains the entity information notebook, which has a tab for each entity type of interest. The default entity types are genes, proteins, and metabolites; new types are easily added by expert users through the explorase R API. Each tab contains a table containing entity information (metadata and analysis results), an expandable panel for filtering the table rows, an expandable panel for hiding or showing table columns and an entry box for searching the table.

The table has a row for each biological entity in the experimental data. The columns of the table contain metadata, such as biological function and biochemical pathway membership. There are two special built-in columns at the left. The first displays the color of the entity chosen by the user. This color matches the color of the glyphs for the entity in the GGobi plots. The other column indicates the user-defined entity lists to which the entity
belongs. The table may be sorted according to a particular column by clicking on the header for the column.

A filtering component, shown in Figure 3, is made visible by clicking on the Filter label above the table in the Entity information panel. This filters the entity information table, as well as the GGobi plots, by any column in the table. Columns containing text data may be filtered according to whether a cell value equals, starts with, ends with, contains, or lacks a user-input text phrase. Regular expression matching (Friedl 2006) is also supported. Numeric values may be tested for being greater than, less than, equal to, or not equal to a user-input number. When filtering by color, the user may choose from the current palette of colors. It is also possible to filter by entity list membership, so that only the entities that belong to a specified list are included in the table. After applying a rule, it may be saved. The saved rules are displayed in a table below the rule editor. There is a checkbox in each row that toggles the activation state of the rule. Buttons allow the deletion of selected rules and the batch activation and deactivation of every rule. Only those cases that pass the intersection of all active filter rules are displayed in the entity information table and the GGobi plots. After filtering, the user might select all the visible entities and save them to an entity list for future recall.

Clicking on the Hide/Show columns label in the Entity information panel expands a component that lists the column names of the entity table and allows the user to specify whether each column is hidden or shown. This helps keep the table clean when generating many columns holding analysis results.

Just above the entity information table is a quick search bar to search individual columns. The user selects the column to search in the pull-down menu on the left and enters a
query in the text box. As the user types, the table scrolls to the first row that matches the query.

Sample/Treatments panel

To the left of the Entity information panel are two panels. The upper one lists the biological samples, such as chips for a microarray experiment, that are in the experimental data. The user may select samples from the list in order to limit the scope of the analysis. It is possible to select a range of samples by holding down the SHIFT key and clicking on the end-points of the range. This is particularly useful after the list has been sorted by an experimental factor using the experimental design table described in the next paragraph.

Clicking on the Details button below the Sample/Treatments list displays a table describing the experimental design, as shown in Figure 4. There is a column for each factor in the experiment, and the rows correspond to conditions. Similar to the Entity information panel, clicking on a column header sorts the table, as well as the Sample/Treatments list in the main GUI, by that column.

Lists/Pathways panel

The bottom panel contains user-defined entity lists. These lists store a group of user-selected entities, usually based on the result of an analysis. Selecting an entity list automatically selects the corresponding entities from the list in the entity information tables. The selected entities may then, for example, be brushed or sent as a query to AtGeneSearch, as described in Section 4.2.

4.2. Toolbar
Above the main GUI panels is the toolbar (Figure 1), which contains many important buttons. We will describe the buttons from left to right. The button with the folder icon provides a shortcut for loading a project. Perhaps the most important button is the brush tool, which appears as a rectangle filled with the current brush color. Clicking on the brush button colors the selected entities in the currently visible entity information table and the same entities in the GGobi plots. The color is selected from a palette that drops down from the button. Besides the brush button are two more brush-related buttons: the first for resetting the colors to the default (gray) and the other for updating the colors in the entity information table to match those in GGobi. The next button to the right, shown in the screenshot as a single curved arrow, queries AtGeneSearch, a web interface to MetNetDB for accessing additional metadata about the selected entities (Wurtele et al. 2003, 2007). AtGeneSearch provides links to other web data sources. The right-most button creates an entity list from the entities selected in the entity information tables and adds the nascent list to the Lists/Pathways panel.

4.3. Menubar

File menu

At the top of the main explorase GUI is the menubar (Figure 1). The File menu provides options for loading and saving files and projects. The Open item is for loading already-created projects. To load individual files and merge them into an opened project the user may choose the Import File(s) item and select the files using a dialog. The Save item saves the entire project to a user-specified directory. The user may select an item from the Export submenu to save an individual project component, such as the entity information or a selected entity list, to a CSV file.
**Analysis menu**

The Analysis menu lists a collection of numerical analysis methods, as described in 3.1. The first set of methods consists of distance measures for comparing the levels of each entity between two samples or conditions. The next set of methods are distance measures for comparing a selected entity against the rest. The final two methods are hierarchical clustering and pattern finding. The cluster results are displayed in an interactive R plot, shown in Figure 5; clicking on a branch point of this R plot brushes the descendent entities. The results of each analysis are added as a column in the entity information table and as a variable in GGobi. This allows the user to sort and filter according to the results of the statistical analyses, as well as visualize these results in GGobi.

A unique feature of explorase is the pattern finder, which calculates whether an entity is significantly rising or dropping relative to the others for each sample transition. The results are displayed as arrows embedded in the entity information table, as shown in Figure 6. The dialog named Find Patterns allows the user to query for specific patterns. To specify a query pattern, the user chooses whether a particular transition should be “Up”, “Down” or the “Same” (these match the terms used by the Find Patterns dialog). The vertical slider on the left of the dialog specifies the number of entity transitions, centered on the median, that are considered to stay the “Same”. All transitions less than the assumed “Same” transitions are considered “Down” and the remaining transitions are considered “Up”. When the Find button is clicked, the entities with matching patterns are selected in the entity table.

**Modeling menu**
The Modeling menu launches graphical interfaces to linear modeling tools in R. Both interfaces are shown in Figure 7. The limma interface leverages the limma package (Smyth 2005) from Bioconductor (Gentleman et al. 2004). The interface prompts the user for the factors to include in the model, including interactions among factors. The user may also choose which results (p values, corrected p values, F statistics, coefficients, or fitted values) to include in the entity information table and GGobi. An Advanced drop-down offers additional options, such as the method for p value adjustment and tests for time linearity.

An interface is also provided for time-course modeling. It fits a polynomial model in time. The user may define the degree of the time polynomial and choose whether the time variable should be treated as a quantitative variable (actual) or as an ordinal variable (virtual).

Tools menu

The Tools menu contains methods for processing experimental data. There are items for calculating replicate means, medians or standard deviations and adding them to the data. The means and medians are automatically included in the list of experimental conditions, so that they may be used in numerical analysis. The second option launches the dialog shown in Figure 8 that provides several simple rules for filtering out entities based on the experimental data. The cutoffs are based on minimum value, minimum fold-change, and maximum variance between replicates. This helps the user focus on particular aspects of the data, such as entities that are changing more between treatments than within. The user may enter the test values directly or use the slider to get some idea of the range of values.

5. Getting started
The first step towards analyzing data with explorase is to load the data. The explorase package has not been designed for data preprocessing, so all preprocessing must be done before loading data into explorase. Usually this involves steps like normalizing and log transforming the data. All files read by explorase must adhere to the comma separated value (CSV) format, as interpreted by the R CSV parser. This format is compatible with the output of Bioconductor tools and the CSV export utility of Microsoft Excel. Accordingly, the file containing the matrix of experimental measurements must be formatted as CSV, with the values from each sample (i.e., chips in a microarray experiment) stored as a column. The first row should hold the names of the corresponding samples. The first column, which does not require a name in the first row, should hold unique ids for each biological entity (transcript, protein, etc.) measured in the experiment.

In addition to the experimental measurements, explorase supports (and, for some features, requires) several types of metadata, all formatted as CSV. The experimental design matrix is required for linear modeling and aggregating replicates. Like the experimental data, the first row should name the design factors, such as genotype, time, and replicate. Some factor names have special meaning. In particular, time is used as a factor in the temporal modeling tool and replicate is used in linear modeling and averaging over replicates. Each cell in the first column of the design matrix should match one of the sample names in the experimental data.

Another type of metadata is the entity annotations that are shown in the central table of the explorase GUI. The only restriction is that the first column should hold entity identifiers that match those of the experimental data.
Finally, entity lists are stored as one or two column matrices. If two columns are present, the first column is interpreted as the type of the entity, such as gene, prot, or met. This allows storing entities of different types in the same list. The other column holds the identifiers of the entities that belong to the list. The name of that column is the name of the list in the explorase GUI.

In order to automatically detect the type of data being loaded, explorase expects the input files to be named according to a specific convention. The mapping from data type to filename extension is given in Table 1. The user must ensure that the input files are named according to that convention.

The data loading process is further simplified by support for projects: all of the data files may be placed into an empty folder and loaded in a single step by choosing the folder in the open project dialog. The types of the files are determined by their file extension.

6. Demonstration

In order to briefly demonstrate the features of explorase, we consider a microarray dataset from an experiment investigating the response of biotin-deficient Arabidopsis mutants to treatment with exogenous biotin (Cook et al. 2007). The mutants were analyzed with and without biotin treatment. Wildtype plants were used as a control and there were two replicates for each set of conditions. Figure 4 summarizes the experimental design. The dataset was normalized using the RMA method.

The first step, after launching explorase, is to load the data. One easy way to load data into explorase is as a project. Projects are directories in the file system that contain the experimental data, design matrix, entity metadata, entity lists, etc, as files. A zip archive
containing an example explorase project for the biotin data, with correctly formatted files, is
provided on the explorase website (Lawrence 2007).

To load the project:

1. Click the Open button at the left-end of the toolbar (see Figure 1).
2. In the file open dialog, select the biotin directory from the (uncompressed) zip archive
   and click Open.

The primary goal of this example analysis is to determine which genes appear to respond to biotin treatment in the mutant. In order to compare across conditions without having to consider each replicate individually, the replicate mean values should be added to the experimental data, assuming that there are no major inconsistencies within the replicate pairs. To add the means to the data: choose the Average over the replicates option from the Tools menu.

Figure 9 displays the result of subtracting the untreated mutant mean from the treated mutant mean in the sample dataset. Sorting by the difference column in the information table allows the coloring of the selected extreme rows using the explorase brush button. Alternatively, the user could brush the outlying points in the GGobi plots and then update the colors in the entity information table to match those in GGobi. The genes at each extreme are grouped into entity lists. The genes brushed in pink are those that have higher expression in the untreated plants compared to the treated, while the blue have lower expression in untreated plants.

To color and group the entities with the most extreme differences between treated and untreated mutant means, follow these steps (as illustrated in Figure 9):
1. Select bio1.no.mean and bio1.yes.mean in Samples/Treatments panel (use the CTRL key for multiple selections).

2. Open the Analysis menu in the menubar at the top of Figure 1. Choose the Subtract item from the Find Difference (two conditions) submenu.

3. Once the column containing the differences appears in the entity information table (the large table in the center of Figure 1), click on the column header (diff.bio1...) until the results are sorted in decreasing order.

4. Select a range of rows at the top of the entity information table (i.e., by holding down the SHIFT key).

5. Click on the downward-pointing arrow on the right side of the Brush button and select the blue color.

6. Click the Brush button to color the selected rows blue.

7. Click the Create List button in the toolbar and enter “treated-high” into the entry that appears in the Lists/Pathways panel at the bottom-left of Figure 1.

8. Click on the header of the difference column of the entity information table again to resort the rows of the entity table so that the rows are in increasing order.

9. Repeat steps 4-7 but use pink rather than blue as the brush color and enter “untreated-high” when creating the list.

10. To sort the rows in the entity table by their list membership, click on the header of the List column in the entity information table.

The GGobi scatterplot at the top-right of Figure 9 compares the two means, showing that the colored observations are indeed outliers. Below the scatterplot is a histogram showing the distribution of the difference.
To create GGobi plots follow these steps (as illustrated in Figure 10):

1. Use the GGobi control panel window (shown in Figure 10): select the New Scatterplot Display option from the Displays menu.
2. Select the two mean variables by clicking on the X button next to the bio1.no.mean label and the Y button next to the bio1.yes.mean label.
3. Select the New Scatterplot Display option from the Displays menu.
4. To change the scatterplot to an ASH plot (histogram), select the 1D Plot from the View menu.
5. Click the X button next to the diff.bio1... variable, so that the histogram shows the distribution of the differences.

One way to verify that those genes are indeed dependent on biotin treatment would be to fit linear models using limma, including effects for the genotype, biotin treatment, and their interaction.

To do this, use the limma interface in explorase:

1. Choose the Linear modeling (limma) option from the Modeling menu (in the menubar in Figure 1). This should open the Linear modeling via Limma dialog (shown on the left in Figure 7).
2. From the list of factors in the Linear modeling via Limma dialog, select the checkbox for the interaction of genotype and biotin. Note that this automatically selects the individual factors.
3. Click the Apply button to run limma.
Figure 11 shows the F values for the interaction of biotin and genotype. The table is sorted by the F value and filtered so that only the genes with the largest F values (F > 95) are included in the table. As one might expect, several of the pink-and blue-colored genes have extreme F values, indicating that biotin treatment has a genotype-dependent effect on those genes.

To identify those genes with the largest F values, follow these steps (as illustrated in Figure 11):

1. Click on the Filter label above the entity information table, so that the filter GUI is shown.
2. Select F.genotype*biotin from the left-most drop-down menu in the filter panel.
3. Change the second combo box to >.
4. Enter “95” into the text field to the right.
5. Click Apply to apply the filter rule (F.genotype*biotin > 95). Genes with an F value less than 95 are now excluded from the entity table.
6. Click the header of the F.genotype*biotin column to sort by it.

One outlier is easy to recognize even from the table: 13212_s.at. The annotations in the table describe 13212_s.at as a glycosyl hydrolase. The functions of the other outlying genes, if known, may be found in the table, and if more information is needed, clicking the AtGeneSearch button in the toolbar spawns a web browser and queries the MetNetDB (Wurtele et al. 2003) for additional details.

This analysis could continue along many paths. For example, the user might search for genes that are similar to the 13212_s.at using the distance measures in explorase (from the Analysis menu), or could continue to inspect the output of limma using GGobi graphics.
Lists of genes could be exported to MetNet for pathway display (MetNet/Cytoscape) or evaluation in the context of public microarray data (MetaOmGraph) (Wurtele et al. 2007). This example demonstrates only a fraction of the potential of explorase.

7. Technical considerations

7.1. Suggested limit on number of samples

There are several reasons behind the suggestion that the number of samples (columns) in the experimental data be limited to a relatively small number (approximately 50 in our experience).

- The explorase and GGobi GUIs are not designed for a large number of variables. In particular, it is cumbersome to navigate through and operate on the variable selection panels.
- The analytical methods in explorase have not been designed/selected with a large number of samples in mind, though many of the same methods would still apply.
- The explorase implementation (and, in general, R itself) has not been heavily optimized (in terms of space and time) for large numbers of samples nor extremely large data matrices.

The explorase package could be modified to handle experiments with large numbers of samples, but it may be more feasible to develop a separate tool for that purpose.

7.2. Software infrastructure

explorase is written in the R language, facilitating integration with R analysis packages. This also enables other R packages to integrate with explorase via its public API that is documented through the online R help in the explorase package.
In order to provide its GUI, explorase relies on the RGtk2 package (Lawrence and Temple Lang 2008), a bridge from R to the GTK+ 2.0 cross-platform widget library (The GTK+ Team 2008). RGtk2 allows explorase to present, completely from within R, a visually pleasing, featureful GUI that is identical across all major computing platforms.

GGobi serves as the visualization component of explorase. The rggobi package (Temple Lang and Swayne 2001; Temple Lang et al. 2008) links R with GGobi. With rggobi, R packages are able to load data from R into GGobi, retrieve GGobi datasets into R, get and set the color of observations, create and configure displays, and more. explorase uses rggobi to load high-throughput datasets and synchronize the color of observations in GGobi plots with the colors in the entity information table in the explorase GUI. This provides the key visual link between the GUI of explorase and the visualizations of GGobi.

8. Related work

explorase is unique among open-source tools in its integration of interactive graphics with R statistical analysis beneath a GUI designed especially for the biologist. The commercial microarray analysis program GeneSpring links to R and Bioconductor but offers limited interactive graphics. The free program Cytoscape (Shannon et al. 2003) is designed for viewing and analyzing experimental data in the context of biological networks and is integrated with R via plugins. However, it lacks interactive graphics outside of its network diagrams.

Many GUIs have been constructed in R, including several in Bioconductor. The limmaGUI package (Smyth 2005) provides a GUI that leads the user from preprocessing microarray data to modeling it with limma and producing reports. Unfortunately, limmaGUI lacks the interactive graphics and breadth of analysis features of explorase. The Bioconductor
iSPlot (Whalen 2005) package provides general interactive graphics using the R graphics engine but offers only a small subset of GGobi’s functionality. Rattle (Williams 2008) is an RGtk2-based GUI that leverages R as it guides the user through a wide range of data mining tasks.

9. Conclusion

The explorase package is an effective tool for analyzing and visualizing high-throughput biological data. Its direct access to R analysis packages, such as limma and others from the Bioconductor project, allow it to take advantage of the latest advances in statistical methods for bioinformatics. The integration with GGobi, including synchronized brushing and the ability to add analysis results as GGobi variables, empowers explorase to display a wide range of interactive multivariate graphics. All of these advanced statistical features are enveloped within a simplified GUI that is tuned for a biologist.

explorase has not yet realized its full potential. There are three predominant directions of planned improvement. The network visualizations in GGobi will be integrated with a convenient and reliable means for matching the identifiers of experimental measurements and nodes in biological networks. This will automatically match identifiers and provide diagnostics to help the biologist ensure the fidelity of the matchings. Analysis methods will be expanded, with a particular focus on clustering and metabolomic analysis. Finally, the visualization of categorical data, such as GO terms and cluster assignments, will be enhanced.

Acknowledgments

We thank Suh-Yeon Choi, Ling Li and others in the MetNet group at Iowa State University for their helpful feedback in the development of explorase. We also gratefully acknowledge our funding sources, NSF Arabidopsis 2010 DB10209809 and DB10520267.
References


Figure 1: The main GUI of explorase with the three main panels denoted by the yellow blurbs. The Entity information panel, located on the right in the GUI, is dominated by the entity information table, which contains metadata and analysis results for each entity in the experiment. Above the table is a quick search bar and tools for filtering the rows and columns of the table. To the left of the Entity information panel are two panels. The top one is the Samples/Treatments panel and lists the samples, such as microarray chips, in an experiment. Below is the Lists/Pathways panel, which contains the names of user-created entity lists. At the top of the GUI are the toolbar and menubar with options for loading and saving the data, analyzing the data and performing other functions. Please note that the appearance of the GUI (especially the button icons) depends on the GTK+ theme, so there may be some aesthetic differences between the screenshots in this paper and a user installation of explorase.
Figure 2: Visualizing and validating Limma results with explorase. The dataset is from the microarray experiment described in Section 6 (Cook et al. 2007). On the right is a scatterplot of p value versus coefficient for the genotype contrast fit by Limma. By highlighting a point in the scatterplot, the corresponding profile in the parallel coordinate plot on the left is also highlighted.
Figure 3: The explorase filter GUI that expands above the entity information table. The active filter rule accepts only the genes with a F genotype* biotin value greater than 120.

Figure 4: The experimental design table, with a column for each factor and a row for each condition.
Figure 5: The hierarchical cluster browser. Clicking on a node selects its children in the explorase information table. In this screenshot, the “treated-high” genes (originally all colored blue) have been clustered according to the mutant chips. The left child of the root has been clicked, resulting in the coloring of five genes in explorase (and GGobi) with the current brush color (red). It appears from the parallel coordinate plot that the patterns of the red genes have more in common with each other than with the blue genes.
**Figure 6:** The pattern finder. The calculated patterns across a time course experiment are shown in the “pattern...” column as arrows representing the direction of each transition. The pattern finder dialog selects rows that match the specified pattern. Here, a constantly increasing pattern has been selected. The vertical slider on the left of the dialog specifies the number of entity transitions, centered on the median, that are considered to stay the “Same”. All transitions less than the assumed “Same” transitions are considered “Down” and the remaining transitions are considered “Up”.
Figure 7: Linear modeling dialogs. On the left is the interface to limma, and on the right is the temporal modeling interface. The user may select the factors and outputs of interest, as well as specify various other parameters.

Figure 8: Simple subsetting GUI. The user may activate rules that filter entities based on their level, fold change, and replicate variance.
<table>
<thead>
<tr>
<th>Type</th>
<th>File Extension (+ .csv)</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metabolomic Data Metabolite Information Metabolite Exp. Design</td>
<td>met.data met.info met.design</td>
<td>suh-yeon.met.data.csv suh-yeon.met.info.csv suh-yeon.met.design.csv</td>
</tr>
<tr>
<td>Proteomic Data Protein Information Protein Exp. Design</td>
<td>prot.data prot.info prot.design</td>
<td>some-proteins.prot.data.csv some-proteins.prot.info.csv some-proteins.prot.design.csv</td>
</tr>
<tr>
<td>Interesting Entities</td>
<td>list</td>
<td>favorite-metabolites.list.csv</td>
</tr>
</tbody>
</table>

**Table 1:** Mapping from data type to filename extension per the explorase filenaming convention. explorase requires project files to be named using these file extensions. An example project with correctly formatted files is in the supplemental data and available from the explorase website (Lawrence 2007).
Figure 9: Difference calculation between the biotin mutant with and without external biotin.

The GGobi scatterplot (above) compares the two conditions; below is a histogram of the difference calculation. The numbered red circles drawn on top of the screenshot illustrate the steps to calculate the differences, as explained in the example.
Figure 10: The GGobi control panel, used for creating and configuring GGobi plots. The File menu supports loading and saving of GGobi datasets, but the explorase user is expected to load and save data through the explorase GUI. The Displays menu contains items for opening each type of display, such as scatterplots and parallel coordinate plots. The View menu allows toggling between display view modes, such as histogram vs. XY plot. The Interaction menu has an item for activating each interaction mode, such as the brush. The Tools menu contains various utilities. Below the menubar are two panes. The left contains options for the current interaction mode. The other lists the variables in the current dataset and has toggle buttons for specifying which variables are plotted in the current display. The numbered red circles drawn on top of the screenshot illustrate the steps to create the plots mentioned in the example.
Figure 11: limma results for the biotin data. Only the entities with an F statistic > 95 for the interaction of genotype and biotin are displayed. The numbered red circles drawn on top of the screenshot illustrate the steps to filter the table.
APPENDIX B. METNETDB: A PLANT BIOLOGICAL NETWORK DATABASE BASED ON A LABELED GRAPH MODEL

A paper to be published in Plant Physiology.

Jie Li\textsuperscript{1}, Leslie Miller\textsuperscript{1}, Heather L. Babka\textsuperscript{2}, Nick Ransom\textsuperscript{2}, Yves Sucaet\textsuperscript{2}, Julie A. Dickerson\textsuperscript{3}, Eve Syrkin Wurtele\textsuperscript{2}

ABSTRACT

Capturing the features of biological networks provides a major challenge in part because of the complex interrelationships and the many uncertainties and incompleteness inherent in the data. MetNetDB is a metabolic and regulatory network database for Arabidopsis. The data is stored as an integrated labeled graph model. This labeled graph model facilitates combining known and hypothetical interrelationships and annotation from multiple data sources.

Aspects of biological networks stored in the MetNetDB labeled graph model include:

- biological entities (e.g., DNA, RNA, polypeptide, protein complex, or metabolite) annotated by subcellular localization, synonyms, literature citations, gene annotations, and/or metabolite formula; in addition, a hierarchy of interaction types (broad categories include catalysis, conversion, transcription, regulation) describe the relationships among the entities.

To facilitate network analysis, biological entities and interactions are represented as nodes, and the associations between them are represented as edges. Properties of entities and interactions are stored as corresponding node labels. Stoichiometric and kinetic parameters

\textsuperscript{1} Computer Science; Iowa State University; Ames, IA 50011

\textsuperscript{2} Department of Genetics, Development, and Cell Biology; Iowa State University; Ames, IA 50011

\textsuperscript{3} Electrical and Computer Engineering; Iowa State University; Ames, IA 50011
for interactions are captured as edge labels. A curator tool supports searching, visualizing and curating the network. The history of each change is retained. These novel features are crucial to enable both integration and storage of the network combined with graph-based search and analysis of the network structures. The network can be shared through Cytoscape-compatible XML files, SBML files, or via a dedicated API. MetNetDB serves as the primary data repository for the MetNet suite of visualization and analysis tools.

Availability: http://www.metnetdb.org/MetNet_db.htm

INTRODUCTION

Predictive biology, one of the great challenges of the current century, requires the power of bioinformatics. Databases can provide a platform for integration and viewing of complex biological networks, as well as for combining high-throughput transcriptomic, proteomic and metabolomic data together with the networks they are associated with. In addition, databases can be designed to enable computational analysis of the biological networks, and for modeling experimental data in the context of known and/or hypothesized network(s). Such analyses facilitate the development of experimentally-testable hypotheses concerning the functions of biological molecules, metabolic or regulatory interactions, or network structures.

Unlike sequence or molecular structure databases, databases for biological networks store data that contain complex internal relationships. For example, there are multiple types of biological entities; these include DNA, RNA, polypeptides, proteins, and metabolites as well as hybrids of these, such as acetylated histones, methylated DNA, proteoglycans, glycoproteins and glycolipoproteins. Furthermore, many types of interactions can occur between these entities (e.g., catalysis, transport, complex formation, allosteric inhibition, transcriptional regulation). These interaction types have a wide range of kinetic and
stoichiometric parameters, thus significant implications for modeling the network. Adding to the inherent complexity in biological networks, there is a tremendous amount of uncertain data and missing data. Even in Arabidopsis, the precise functions of most genes are not yet understood (Swarbreck et al., 2008), regulatory interactions are even less comprehensively understood, and kinetic information is rare (Oliver et al., 2002).

To best facilitate visualization and analysis, a biological network database needs to be able to capture and represent biological interrelationships in many ways. To accomplish this, the data storage model must be rich enough to describe both the complex relationships and the uncertainties inherent in the network. To take advantage of the continuously expanding battery of experimental technologies that give increasingly detailed information about molecular and spatial relationships (e.g., ChIP-hybrid interactions; Wu et al., 2006; Bindila and Peter-Katalinić, 2009); protein localization (Sadowski et al., 2008); laser desorption-GC-MS (Chen, 2008)), the model needs to be flexible enough to incorporate not only the current types of biological data, but also to accommodate future data types. The data model should also be able to store, manipulate and export biological network data easily, and to incorporate new information from existing data sources effortlessly.

Biological network databases have been implemented as object, frame, or graph data models (Table I). The object data model is based on object-oriented programming principles, and stores data as a collection of objects (Booch, 1994). Biological concepts are described through class definitions and individual instances are represented as objects (e.g., molecules, interactions, pathways, networks). The object data model suffers because there is minimal connectivity between objects, so the network cannot be easily modeled. The frame data model (Reimer and Hahn, 1983; Karp and Paley, 1996) is a collection of frames that
represent knowledge. Frames can be either a class (e.g., a concept like “polypeptide”) or an instance (e.g., a concrete biological entity like “acetyl-CoA”). The frame model uses slots to define the relationships among frames. However, the frame model, similar to the object data model, cannot be used easily for predictions about the behavior of the network. Graph models are based on nodes and their interconnecting edges (Cormen et al., 2003). Graph models have an inherent advantage over object and frame data models: a graph implicitly represents a network. (A network can be represented by object or frame data models artificially, at best). Thus, graph models provide a natural approach for representation and manipulation of biological networks.

Graph-based approaches to represent complex data content have been studied for decades (Renzo and Claudio, 2008). Several graph models have been developed for computational analysis of biological and chemical phenomena, including graphs of chemical structures, chemical reaction graphs, bipartite graphs, and hypergraphs (Deville et al., 2003). Each of these graph models is designed for a specific type of computational analysis. The simple graph model is the basis of all other types of graph models. It contains nodes and edges that can carry limited information; the edges can be directed or undirected. Many existing graph algorithms are developed based on the simple graph model (Cormen et al., 2003). Chemical compound graphs, reaction graphs, bipartite graphs and hypergraphs provide a slightly richer structure to highlight specific information such as chemical structures, reactions or topology relationships for a specific analysis. However, these representations are either incomplete or ambiguous for complex network storage (Deville et al., 2003). The hypernode graph model used in PATIKA (Demir et al., 2002, 2004; Dogrusoz et al., 2006) allows the nodes to represent graphs, thus creating a nested graph. This model is designed to incorporate multiple
levels of abstraction for pathways and to represent state transitions of molecular complexes. Incomplete pathway knowledge can be stored in this model. However, the hypernode graph model in addition capabilities for checking data integrity are limited (Renzo and Claudio, 2008). SAGA (Tian et al., 2007) uses an attribute graph model, in which the nodes contain attributes that can be used in an index-based graph alignment algorithm. Although this representation is powerful enough to be searched, it cannot store complete information on the stoichiometry or kinetic parameters of each interaction. While these existing graph models are good for different aspects of graph-based network analysis, none is designed to be able to store many different types of data.

In this paper, we introduce the MetNetDB labeled graph model and present an overview of the MetNetDB system. The MetNetDB model is designed to store diverse information about the biological networks. It represents both molecular entities and interactions as nodes (subtypes are used to distinguish molecule and interaction types). It also defines the graph operations and rules for data integrity. Meanwhile, this model can be used flexibly for graph-based analysis. For example, it supports the ad-hoc reconstruction of subnetworks of interconnecting entities through a p-neighborhood search. The labeled graph model has been implemented through MetNetDB, which is the central data repository of the MetNet project (Wurtele et al., 2003, 2007). MetNetDB stores biological networks and annotations for Arabidopsis and other species. There are four main contributions to this work.

First, the labeled graph model enables many different types of data to be stored and retrieved, while maintaining the ability for graph-based analysis. The model provides flexibility for database expansion and internal or third-party tool implementation.
Second, MetNetDB collects and integrates biological interaction networks and annotations from various data sources. These include metabolic networks, regulatory networks, protein subcellular localizations, and annotations of genes, proteins and chemical compounds.

Third, a curator tool is implemented on top of the labeled graph model. The curator tool allows the curator to create new networks, to modify and update existing data, and to compare pathways from different sources.

Finally, MetNetDB database provides a rich source of integrated network data that can be used in a variety of applications, based on its labeled graph model. For instance, the MetNet tools (Dickerson et al., 2003; Wurtele et al., 2003, 2007; Lee et al., 2004; Ding et al., 2005; Yang et al., 2005; Lawrence et al., 2008; Mentzen et al., 2008; Mentzen and Wurtele, 2008) can access the network data based on the labeled graph model. The network data can also be exported as formatted data files, for example, XML or SBML (Hucka et al., 2003) files, depending on the needs of the biologists and/or tool developers. For Java and .NET application developers, a dedicated development library and an application programming interface (API) are available for download through http://metnet3.vrac.iastate.edu/api.

RESULTS AND DISCUSSION

To capture a complex regulatory and metabolic network requires a data model capable of storing large amounts of network and annotation data. The model also has to be rich enough to incorporate the multidimensional facets of the data. If external data sources are used, it must be able to interpret the varied data models employed by these online data sources, form a superset of interactions from the data in them, and integrate this data with the rest of the MetNetDB biological network (see supplemental data Table S1). In addition, to provide
support for the evolving requirements of systems biology research, the model must be easily extensible. Here, the MetNetDB labeled graph data model and its application to the MetNetDB database are described.

**The MetNetDB Labeled Graph Model**

A labeled graph model can combine the advantages of efficient network analysis, high capacity data storage, representation of multiple data types, and extensibility. Because it is a graph model, it can be analyzed directly by graph algorithms. In addition, the node and edge labels can be used to store large amounts of information on individual biological entities and reactions in tables as flexibly as either the object data model or the frame model, since any biological property can be described as part of a label (attribute, value). In each label, the attribute element indicates the property name, while the value element can hold any type of data. Many types of data can be represented and stored. For example, data for a label can even be a dynamic behavior (for example, MetNetDB could store data on changes in rates of the activity of a protein complex in the presence of increasing levels of an allosteric activator); indeed, any behavior that can be described by SBML or MathML (http://www.w3.org/Math/) can be represented. Therefore, the labeled graph model provides a way to incorporate multiple types of static and dynamic information into the graph structure. The labeled graph model is easily extensible, for example, high throughput data sets from a new type of experimental protocol can be encoded as an XML format string and stored in a new label. Thus, the underlying database only needs to support string value storage. The database schema itself does not need to be changed to adapt to representing new types of data and models.
Technical Representation of MetNetDB

MetNetDB labeled graph model is defined by a 7-tuple $G = (V, E, L, f_v, f_e, O, R)$ in which $V$ is the node set, $E$ is the edge set and $L$ is the set of node and edge labels. Labels have the form (attribute, value). The two functions, $f_v$ and $f_e$, are the node label assignment function ($f_v : V \rightarrow L$) and edge label assignment function ($f_e : E \rightarrow L$), respectively. $O$ is the set of operations defined on the graph model. The edit operations can be node insertion, node deletion, label substitution of nodes and edges, edge insertion, and edge deletion. These five operations allow us to change the status of the graph. $R$ is the set of rules that define data integrity. The rules are obtained from either biological knowledge or graph requirements. Rules are defined as a set of Boolean functions mapping from nodes, edges, and labels to a Boolean domain, e.g., $r : V \times E \times L \rightarrow \{0,1\}$. The rules determine which combination of nodes, edges and labels are allowed or prohibited.

The motivation for this use of graph labels is to combine straightforward graph-based analysis of the network, to permit large amounts of data storage, and to provide a flexible framework that can be extended as needed. Existing forms were considered, but did not satisfy all these criteria. Hypergraphs, for example, can contain multiple types of information. However, they are inherently complex and require transformation to simple graphs to make use of most existing graph algorithms. To illustrate the model, we use the final reaction in the pathway “ethylene biosynthesis from methionine” (Figure 1) to shows an example of the labeled graph model instance for the pathway fragment.
Representing Biological Networks Using MetNetDB Labeled Graph Model

Table II defines some of the terminology used in MetNetDB, and Table III summarizes how MetNetDB implements the model in the database. An “entitywithlocation”, which represents a biomolecule and its subcellular location, is mapped to a node in the labeled graph, and all of its properties are assigned to the label of that node. An interaction is also represented as a node in the model. Edges represent the relationship between an entitywithlocation and an interaction. In a case of catalysis in an enzymatic reaction, a catalysis interaction node is connected to the enzymatic reaction node. The biological properties of each interaction, for example reversibility or strength, are stored in the corresponding node label. The coefficients are stored in the label of the edge between the entitywithlocation and the interaction. The direction of each edge indicates the direction of that interaction. To illustrate the model, we use a metabolic reaction, ethylene biosynthesis from aminocyclopropane carboxylic acid (Figure 1).

Data Content in MetNetDB

MetNetDB provides a wide variety of data to support the study of systems biology in plants: such as metabolic pathways, transcriptional regulatory networks, gene annotations, protein localization information, and metabolite annotations. Figure 6 indicate which external databases are integrated into MetNet. The method of integration is discussed further in the “MetNetDB Implementation” section.

Metabolic and Regulatory Interaction Networks

MetNEtDB currently obtains interaction information from three sources. MetNetDB integrates each release of the metabolic pathway database AraCyc (Zhang et al., 2005),
including metabolic pathways and unique genes assigned to these pathways, annotations and references for genes, metabolites, enzymes and reactions. Data from the transcriptional network database AGRIS (Palaniswamy et al., 2006) is integrated. AGRIS contains information on: binding site of promoters, loci of associated transcriptional factors, and corresponding references; these data are integrated into the MetNetDB network using the interaction types: direct transcriptional activation and direct transcriptional inhibition (Figure 2). Finally, the MetNetDB curator adds signal transduction networks based on the current literature (such networks are not otherwise available as a web resource for Arabidopsis).

Annotation of Genes and Proteins

MetNetDB contains gene annotation associated with the interactions and entities, such as function annotation and external database ID mapping. To expand the annotation, we integrate the full copy of Gene Ontology (Harris et al., 2004) and TAIR gene annotations (Swarbreck et al., 2008). The mapping between the locus name and probe sets of Affymetrix ATH1 and AG genome arrays is included. The UniProt IDs (The UniProt Consortium, 2009) of gene products are associated with the corresponding Arabidopsis loci. MetNetDB also integrates MapMan (Thimm et al., 2004; Usadel et al., 2005) bin annotations, which include gene annotations and functional categories for the gene products (see supplemental data Figure S1). The membership of each gene locus in a regulon (co-expression cluster) is also annotated (Mentzen and Wurtele, 2008).
Annotation of Metabolites

Metabolite information uses the expert-curated annotation from Plant Metabolomics (plantmetabolomics.org), which assigns annotation to metabolites that are detected in the metabolomics analysis by this National Science Foundation Arabidopsis 2010 consortium. For those metabolites not yet assigned by Plant Metabolomics, MetNetDB computationally maps metabolite names to data entries in PubChem (Wheeler et al., 2007). PubChem contains comprehensive information including chemical structures, CAS registry numbers (Buntrock, 2001) from NCI chemical compound database linked to their literature references (Sitzmann et al., 2008), and information from ChEBI, an ontological classification for chemical compounds (de Matos et al., 2006). There are several challenges inherent with this approach due to inconsistencies within some of these databases. In the PubChem database, a CAS registry number often is listed as a synonym, but is not explicitly indicated to be a CAS registry number. In this case, the CAS registry verification rules (David, 1997) are used by MetNetDB to search PubChem, identify those synonyms that are similar to CAS registry numbers, and assign putative CAS numbers to the metabolites. Because of the ambiguities associated with chemical names, MetNetDB provides possible mappings between metabolites in MetNetDB and those in ChEBI by using an exact name match approach. These mappings can then be used by a human curator. Figure S2 in supplemental data illustrates the annotation of the simple metabolite ethylene in MetNetDB, and Figure S3 shows an example for this name mapping process.
Interaction Types

To capture a wide variety of possible biological interactions, MetNetDB currently defines 39 interaction types, including enzymatic reaction, catalysis, transportation (7 types), transcription, translation, positive and negative regulation (9 types for each), and composition relation. The interaction types are derived from GO with several modifications. A full list can be viewed in the MetNetDB curator tool (see supplemental data Table S3). To integrate metabolic regulatory networks, MetNetDB uses transcription/translation interactions, RNAs and polypeptides.

**AND-OR Relationships are a** special group of interaction types dealing with the relationship among polypeptides and protein complexes. The relationship between proteins, their constituent polypeptides, and the genes that encode them is crucial for modeling biological pathways. This is because the composition of a given protein complex determines its activity and function. Composition can vary in several ways. First, protein complexes often can contain products from one or more members of a multigene family. For example, Arabidopsis isoamylase contains one or more polypeptides encoded by the 4-member isoamylase family, and the function of this enzyme, depending on its constituents, is thought to include both starch synthesis and starch degradation (Li et al., 2007; Wattebled et al., 2008). Second, polypeptide components of proteins may be modified covalently. For example, histones can be modified by acetylation, which affects gene expression.

To efficiently represent the relationships among genes and an enzyme in MetNetDB, “AND” and “OR” are used as the operators of the composition relation (Figure 3). “AND” means each of the polypeptides or protein complexes that are needed to form a new protein
complex, while “OR” means several polypeptides can perform a similar function. Many
network databases do not include the protein-gene relationships in the networks directly and
only list the genes associated with the proteins as the annotation. Only a few databases
attempt to describe the composition of protein complexes, using different methods to
describe these relations. For example, Reactome (Matthews et al., 2009) which focuses on
human and vertebrate networks, uses “defined set” to describe the complexes and protein
paralogs; this is a variant of the hypernode graph model concept (Renzo and Claudio, 2008).
Compared to this multilevel/hypergraph graph data model, MetNetDB “flattens” hypernodes
into pathway graphs by introducing the AND-OR relations as a parallel concept of
biochemical interactions. This simplifies the application of many graph algorithms in the
pathway analysis without losing biological information.

Representation of protein complexes in a database suffers from the current paucity of
knowledge of which protein complexes actually exist, and of the effect of each particular
combination of polypeptides on protein activity/function.

Annotation of Subcellular Localization

Interconnectivity among cellular compartments is critical. However, many biomolecules cannot pass
across compartmental barriers. Therefore, subcellular information is critical to understanding and
modeling network flow. In addition, many reactions (e.g., carboxylation of acetyl-CoA in the cytosol
and in the plastid, Nikolau et al., 2003) and even entire pathways (e.g., cytosolic and plastidic
glycolysis, Pichersky and Gottlieb, 1984; Givan, 1999) are present in more than one compartment and
thus are modulated by the distinct environment of the particular compartment in which they reside.
Currently, there is a moderate knowledge of the location of many of the better-understood cellular
interactions. To capture this knowledge, MetNetDB incorporates subcellular localization information
from a combination of sources. The MetNet curator assigns location information to proteins involved in MetNet pathways using original literature references and communications from expert researchers. In addition, information is incorporated from PPDB (Friso et al., 2004; Sun et al., 2009) and plprot (Kleffmann et al., 2006), which contain plastid proteins of Arabidopsis and other plants; AMPDB (Heazlewood and Millar, 2005), which contains Arabidopsis mitochondrial protein data; AraPerox (Reumann et al., 2004), which contains peroxisomal Arabidopsis proteins; and AtNoPDB (Brown et al., 2005), which contains Arabidopsis nucleolar proteins. These protein localization databases include annotation from a combination of experimentation, literature, and predictions. If there is no experimentally-derived information for a protein, locations are assigned computationally from the protein sequence based on sequence similarity using TargetP (Emanuelsson et al., 2000). This computational method is superseded if there is experimental information from protein localization databases.

Each location annotation includes evidence and references/sources, to help users evaluate whether the location assignment for a given protein is based on a sequence-based prediction, a single experiment, or multiple experiments. The external data source is stored in the “comment” field and the evidence is encoded in the “location confidence” field. To accommodate location information, MetNetDB employs a hierarchy of more than 70 subcellular locations (see supplemental data Table S2). This hierarchy is predominately based on GO categorization combined with locations in existing protein localization databases. The subcellular location list can be expanded as required.

**Synonyms and Abbreviations**

Biomolecules often have multiple names/synonyms; conversely, a given name may describe more than one biomolecule. Such inconsistencies cause problems in integrating network data
from different sources. For example, ACC can be used as the abbreviation for acetyl-CoA carboxylase, an enzyme required for de novo fatty acid synthesis, or for 1-aminocyclopropane-1-carboxylic acid, an intermediate in the synthesis of ethylene. Ontologies and controlled vocabularies are increasingly being implemented to reduce this problem (Smith et al., 2007). Despite these continuing efforts, ambiguities remain.

To address the ambiguity of terminology, MetNetDB uses unique IDs from standardized databases and ontologies to reference each biomolecule. Arabidopsis locus IDs are used to identify Arabidopsis genes, RNAs, and polypeptides. UNIQUE-IDs from the BioCyc open chemical compound database are used, when available, to identify metabolites. In parallel, MetNetDB expands lists of synonyms to facilitate data integration. Most of the synonyms currently in MetNetDB are computationally extracted from online databases, such as BioCyc’s open chemical database, KEGG (Kanehisa et al., 2004; Hashimoto et al., 2005), ChEBI, PubChem, TAIR, and BRENDA (Chang et al., 2009). Other synonyms are added computationally, based on common abbreviations (for example, CoA is a synonym for coenzyme A, so acetoacetyl CoA is considered a synonym for acetoacetyl coenzyme A). Still other synonyms are added from the literature or manually extracted from on-line databases by curators.

*Literature References and User Comments*

MetNetDB extracts literature references from several sources, including AraCyc and AGRIS for interactions, ChEBI for chemical compounds, and TAIR for genes and corresponding RNAs and proteins (Figure 4). MetNetDB curators also add references via the MetNetDB
curator tool. This tool integrates with PathBinder (Ding, 2003; Wurtele et al., 2007) to allow rapid addition of references from PubMed.

Users can also input comments and references on pathways, interactions and biological entities. Comments can provide additional annotation of genes, source of data, reliability, or any information that does not fit in existing MetNet fields.

**Implementation of MetNetDB**

**Data Integration**

MetNetDB combines biological interaction network data and annotation data from a wide variety of data sources (Figure 4). All information collected in MetNetDB from external data sources is represented in the MetNetDB labeled graph model. Annotations from different sources are aggregated according to the bimolecular IDs. Hands-on expert curation is required when there is ambiguity or conflict information during the integration.

**Pathway Comparisons**

The curator tool compares two or more pathways graphically to identify differences. This is important for incorporating new pathways into MetNetDB and for reviewing pathway changes. The curator tool supports two types of pathway comparisons. Using an “exact match” mode, it will match all biological properties of biomolecules. Alternately, the curator tool can also match all properties except the subcellular location. The first mode helps a curator to identify incremental modifications of a pathway. The second mode helps a curator to identify the different parts of MetNetDB pathways that have locations and pathways from
other databases, such as AraCyc updates, which often do not contain subcellular location information for biological molecules in pathways (see supplementary data Figure S4).

Similar to other biological network databases, MetNetDB supports keyword-based querying of its database. In addition, MetNetDB supports querying based on a list of user-specified molecules, or a partial-graph structure of a pathway (Figure 5). “Exact matches” and “approximate matches” are supported in these query types. Querying based on a list of user-specified biomolecules lets users map experimental data (e.g., “omics data”) onto the MetNetDB biological network. Such mapping can give scientists insight of their data in a systematic way and generate new hypotheses (Parley and Karp, 2006; Babur et al., 2008; Kincaid et al., 2008). Querying based on a partial structure allows a curator to identify whether part of a pathway to be integrated (the query pathway) is already in MetNetDB. If a portion of the new pathway exists in MetNetDB, the curator can add new interactions to the existing pathway. Figure 5 show an example of pathway querying based on a partial structure. In panel A of Figure 5, the curator draws a simple structure involving three metabolites (L-methionine, S-Adomethionene, and 5-methylthioribose). Panel B of Figure 5 displays the best matching alignment found in MetNetDB.

Querying in MetNetDB is implemented via graph matching based on Messmer’s subgraph isomorphism algorithm (Messmer and Bunke, 1998) and MetNetDB’s subgraph extraction algorithm (Li, 2008). This subgraph extraction algorithm is designed to extract a subgraph from a network database based on a list of nodes such as genes, metabolites, as well as to extract common substructures from the Messmer’s isomorphism results.
Active Rule-checking During Curation

MetNetDB implemented a set of constraints, or rules, that help guarantee data integrity. These rules are classified into two types: biological and graph. Biological rules are to make the biological network comply with conventional biological knowledge. For instance, “an enzyme is not a metabolite” is an example of a biological rule in MetNetDB. Graph rules guarantee the basic data integrity in the context of the graph. For example, a pathway should always be connected (i.e. there is no orphan nodes in the graph). Rules are automatically checked each time a curator submits data or edits a pathway. The rules are classified as mandatory or optional. Any violation of mandatory rules will block the data submission. Those violations will be highlighted with a red box in the curator tool. Conflicts with optional rules allow exceptional types of biological data to be saved and curated but provide a warning to the curator by placing a pink box around the expected errors. For example “an enzyme is a protein or protein complex” is an example of a rule that is designated optional. Usually, it is the case and the rule warns that the curator or external database may have input an error; however, there are important exceptions, for example RNA molecules can also be enzymes. The curator can make the decision. Rules can be easily added or removed. Table S4 in supplemental data displays all rules currently implemented in the MetNetDB curator tool.

Tracking Data Changes and Concurrency Control

An important feature of MetNetDB is that it supports tracking of changes in the database. There is a record of each change that is made. Unlike Meta-All (Weise et al., 2006), which only stores version when users request it, MetNetDB automatically stores every type of change and all the history of these changes (Figure 6). For instance, if an entity with location
is deleted from the database, MetNetDB saves this deletion and any resultant modifications to interactions and pathways. Also, if an interaction or a pathway is modified, the previous version can be retrieved and compared. Finally, removed data can be undeleted, if necessary.

In addition to revealing data provenance, tracking the data history provide a method to ensure concurrency control. This feature is important because MetNetDB and its curator tool are designed as a distributed application, enabling multiple curators to operate on the database at the same time. As a safeguard, the curator tool prevents two curators from updating the same database record at the same time. The workflow shown in Figure 7 illustrates how concurrent editing is eliminated. When a curator retrieves data from the database, the corresponding version information is saved at the client side. Before a curator can update the data, the most recent version information is retrieved from the database again to make sure that nothing has been modified since the curator initially retrieved the data. In this way, MetNetDB guarantees that curators always update the data that they are reviewing.

Data Export

Currently, MetNetDB can export biological network data to MetNet tools in several ways: XML formats, direct table dump, and developmental library. For instance, Cytoscape with MetNet plugins (evolved from FCModeler [Dickerson et al., 2003; Wurtele et al., 2007]) and MetNetVR (Yang et al., 2005) currently use XML files generated by MetNet XML Builder (http://www.metnetdb.org/CytoscapeXML/) for network visualization and analysis. A MetNet developmental library (http://metnet3.vrac.iastate.edu/api) is available for two popular software platforms: Java and .NET. In both environments, a set of classes that refer
to logical MetNetDB concepts are available. A tool for exporting to SBML format and BioPAX format (Luciano and Stevens, 2007) can be accessed at http://metnetdb.org/metnet3.

**CONCLUSION**

MetNetDB enables storage and analysis of combined metabolic and regulatory interaction networks. It contains three components: a database, a curator tool and a set of data-sharing interfaces. These components are built on a labeled graph model. Capitalizing on this labeled graph model, the database can support import, integration, update, edit, and export of complicated networks and associated annotations from heterogeneous data sources. In addition, the labeled graph model enabled us to implement features into the curator tool including tracking the history of data changes, validating data input, and comparing or querying different version of biological networks. The graph model also benefits other bioinformatics tools, since they do not need to handle a relational model of the underlying database system. In this context, MetNetDB provides a comprehensive repository supporting systems biology research.

Furthermore, MetNetDB presents an approach to biological network integration. This approach includes data model transformation and integration, semantic mapping, data transformation, and conflict resolution. The inexact graph-matching implementation is based on the labeled graph model and can be used for network integration when these biological networks are represented as the labeled graph model. A similar approach can be used for diverse scientific applications which require the integration of data sources and to provide a single data view for pathway analysis tools.
ACKNOWLEDGMENTS

We greatly appreciate the provision of data and interactions from Peifeng Zhang, Seung Y. Rhee, Lukas Mueller and the AraCyc team (metabolic pathway data). We thank the experts in specific research areas who have contributed their time and data. AGRIS, Björn Usadel and Mark Stitt (MapMan gene annotations), Peter Hedden (GA signal transduction), Basil Nikolau (cuticular wax), Diane Bassham (autophagy), Ron Mittler (oxidative stress), Martha James (starch metabolism), Yingjun Wang (central metabolism), Jianling Peng (signal transduction pathways), and Mentzen Wieslawa (gene correlation data for inexact pathway query and regulon annotations). Sivakumar Swaminathan (MetNetDB curator) has been exemplary in assigning locations to proteins, and many researchers have generously provided him with specific information in their field of expertise. We are grateful to Li Ling for biological suggestions, and Nick Ransom for computational suggestions. Finally, we thank members of the MetNet group for their valuable input and suggestions.

FOOTNOTES

This project is funded by NSF Arabidopsis 2010 #0520267.

LITERATURE CITED


PATIKA: an integrated visual environment for collaborative construction and  


Creating and modeling metabolic and regulatory networks using text mining and  
fuzzy expert systems. Computational Biology and Genome Informatics: 207-238.

Ding J (2003) PathBinder: a sentence repository of biochemical interactions extracted from  
MEDLINE. Master Thesis. Iowa State University, Ames, IA.

Ding J, Viswanathan K, Berleant D, Hughes L, Wurtele ES, Ashlock D, Dickerson JA,  
Fulmer A, Schnable PS (2005) Using the biological taxonomy to access biological  

PATIKAweb: a Web interface for analyzing biological pathways through advanced  

Ellis LB, Roe D, Wackett LP (2006) The University of Minnesota  
biocatalysis/biodegradation database: the first decade. Nucleic Acids Research 34:  
D517-521.

localization of proteins based on their N-terminal amino acid sequence. Journal  
Molecular Biology 300: 1005-1016.


<table>
<thead>
<tr>
<th>Model</th>
<th>Terminology Subtype:Examples</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Object oriented</strong></td>
<td>Classes, objects</td>
<td>Storage</td>
<td>Minimal connectivity between objects; limited visualization and analysis</td>
</tr>
<tr>
<td>(Booch, 1994);</td>
<td>UM-BDD (Ellis et al., 2006)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Collection of objects</td>
<td>aMAZE (Lemer et al., 2004)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Frame</strong></td>
<td>Frames, slots</td>
<td>Storage; semantic restrictions; some simple</td>
<td>Limited visualization and analysis</td>
</tr>
<tr>
<td>(Reimer and Hahn,</td>
<td>BioCyc (Karp et al., 2005)</td>
<td>relationship queries possible</td>
<td></td>
</tr>
<tr>
<td>1983);</td>
<td>Reactome (Matthews et al.,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Collection of frames</td>
<td>2009)</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Graph</strong></td>
<td>Nodes, edges</td>
<td>Simplified models to analyze topology</td>
<td>Hard to integrate various types of networks; limited storage</td>
</tr>
<tr>
<td>Nodes, edges (Cormen</td>
<td><strong>Chemical structure graph, reaction graph</strong> (Deville et al., 2003)</td>
<td>relationships in chemical compounds or reactions</td>
<td></td>
</tr>
<tr>
<td>et al., 2003)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Attribute graph:</strong></td>
<td>SAGA (Tian et al., 2007)</td>
<td>Supporting storing attributes of biomolecules</td>
<td>Cannot store interaction properties; limited storage</td>
</tr>
<tr>
<td></td>
<td></td>
<td>and interactions in the nodes for graph</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>similarity computation</td>
<td></td>
</tr>
<tr>
<td><strong>Bipartite graph,</strong></td>
<td></td>
<td>Can represent compound nodes and reaction nodes</td>
<td>Cannot represent metabolic networks and regulatory networks at same time without extension of the model; limited storage</td>
</tr>
<tr>
<td><strong>hypergraph</strong></td>
<td></td>
<td>in same graph</td>
<td></td>
</tr>
<tr>
<td>(Deville et al., 2003)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Hypernode graph</strong></td>
<td>PATIKA (Demir et al., 2002; Demir et al., 2004; Dogrusoz et al., 2006)</td>
<td>Represents pathway knowledge in multiple abstract levels; storage Retrieval?</td>
<td>Lack of data integrity facility; Intractable to validate all possible pathways. Retrieval?</td>
</tr>
<tr>
<td><strong>Labeled graph</strong></td>
<td>MetNet(this paper)</td>
<td>Large-scale integration of networks;</td>
<td>Complex implementation. Currently lack of a formal definition of biological</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Computationally sophisticated graph-based analysis; Graph-</td>
<td></td>
</tr>
<tr>
<td>Model</td>
<td>Terminology</td>
<td>Subtype:Examples</td>
<td>Advantages</td>
</tr>
<tr>
<td>-------</td>
<td>-------------</td>
<td>------------------</td>
<td>------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>matching algorithms including inexact network matches; Semantic restrictions are possible; Kinetic and stoichiometric parameters can be incorporated into analysis; High network storage and retrieval capacity</td>
</tr>
</tbody>
</table>
Figure 1. MetNetDB representation of ethylene biosynthesis from 1-aminocyclopropane-1-carboxylic acid. A) ACC oxidase reaction. B) The MetNetDB labeled graph model representation. The reaction contains two interactions, EC 1.14.17.4 and the catalytic activity of ACC oxidase. There are eight entity nodes (n1-n8, ovals) and two types of interaction nodes (rectangles). The label of node n8 “ethylene” is shown in the table. The label “2” on the edge between node n4 and n7 is the stoichiometry of biological entity H₂O in the interaction. In this model, the rule “a catalysis reaction must be connected with an enzyme and an enzymatic reaction” is enforced for the relationship among n4, n5 and n6.
Figure 2 Transcriptional regulatory network with metabolic network. This figure displays a part of the catalase ascorbate glutathione pathway combining data from multiple sources. In this Cytoscape (Shannon et al., 2003) graph using MetNet plugins (Dickerson et al., 2003; Wurtele et al., 2007), yellow nodes represent the biological entities located in the nucleus; light yellow nodes, cytosolic entities; grey nodes, microbody; blue lines, catalyses and enzymatic reactions; green lines, positive regulation; dashed lines, low confidence data. This example shows the integration of AGRIS regulatory networks and existing metabolic networks in MetNetDB. AT4G23810 encodes a transcription factor (TF). When adding this AGRIS derived information, we created a positive transcriptional regulation interaction between the gene product of AT4G23810 and its target loci (AT1G20630, AT4G35090, and AT1G20620).
### Table II. Selected terminology used in MetNetDB

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entity</td>
<td>Biological molecule (e.g., gene, RNA, polypeptide, protein complex, metabolite, cis-element) or environmental condition</td>
</tr>
<tr>
<td>Entitywithlocation</td>
<td>An entity in a particular subcellular location, such as an organelle or a membrane. This concept is especially important for computational analysis of experimental data that contains information about subcellular concentration of entities</td>
</tr>
<tr>
<td>Location</td>
<td>Subcellular location, subcellular compartment (Locations currently in use are given in supplemental data Table S2)</td>
</tr>
<tr>
<td>Interaction</td>
<td>Biochemical interaction, or the relationship among entitywithlocations (Interaction types are shown in supplemental data Table S3)</td>
</tr>
<tr>
<td>Left, from</td>
<td>The left part of an interaction when it is written down. (The reactants, for a chemical reaction.)</td>
</tr>
<tr>
<td>Right, to</td>
<td>The right part of an interaction when it is written down. (The products, for a chemical reaction.)</td>
</tr>
<tr>
<td>Reversible</td>
<td>An interaction that is kinetically bidirectional</td>
</tr>
</tbody>
</table>

### Table III. Organization of biological network data in MetNetDB labeled graph model

<table>
<thead>
<tr>
<th>Components of biological interaction network</th>
<th>Representation in the labeled graph model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entitywithlocation</td>
<td>Node</td>
</tr>
<tr>
<td>Biological properties of entitywithlocation (Name, organism, subcellular location, data origin, entity type, synonyms, abbreviation, references, user comments, external database cross-reference, confidence; for gene loci: Affymetrix probe set IDs, Uniprot ID, AraCyc pathway names, regulon membership, GO terms, TAIR annotation, TAIR reference IDs, MapMan BIN ID, MapMan category, MapMan annotation, TargetP location, gene name and symbol, AGRIS regulation, and AGRIS reference; for metabolites: PubChem CID, CAS, PubChem synonyms, ChEBI ID, ChEBI name, formula, IUPAC ID, SMILES). (Currently, MetNetDB defines 7 entity types and 74 subcellular locations, see Table S2)</td>
<td>Node label</td>
</tr>
<tr>
<td>Interaction</td>
<td>Node</td>
</tr>
<tr>
<td>Biological properties of interaction (Organism, data origin, interaction type, EC number, confidence, reversibility, references and user comments). Relationship between entitywithlocation and interaction</td>
<td>Node label</td>
</tr>
<tr>
<td>Relationship between entities and interactions</td>
<td>Edge between entitywithlocation and interaction node</td>
</tr>
<tr>
<td>Relationship between interactions</td>
<td>Edge label</td>
</tr>
<tr>
<td>Stoichiometric coefficient, kinetic data</td>
<td>Edge between interaction nodes</td>
</tr>
<tr>
<td>Relationship between interactions</td>
<td></td>
</tr>
</tbody>
</table>
Figure 3 Composition relationships of the ATP citrate lyase (ACL) protein complex visualized in Cytoscape with MetNet plugins. The ACL protein complex, located in the cytosol, contains two distinct types of subunits: ACL-A and ACL-B (Fatland et al., 2005); the complex formation between ACL-A and ACL-B is thus represented by a composition-AND relation (orange edges). ACL-A can be formed from any of three polypeptides: ACL-A1, ACL-A2, or ACL-A3. Thus, a composition-OR is used to connect these polypeptides with ACL-A (ochre edges). Similarly, a composition-OR is used to connect ACL-B1 and ACL-B2 polypeptides and ACL-B. Off-white nodes indicate entities in the cytosol. Yellow nodes indicate entities are in the nucleus. Node shapes indicate the types of entity: diamond nodes, polypeptides or protein complexes; rectangular nodes, RNAs; hexagonal nodes, genes. Edge colors indicate the interaction types: black edges, transcriptions; pink edges, translations.
Figure 4. Subcellular localization information for the pyruvate dehydrogenase E1 alpha encoded by AT1G01090 visualized in AtGeneSearch (http://metnet.vrac.iastate.edu/MetNet_atGeneSearch.htm). There are two data sources for localization: experimental data, in this case, the mitochondrial protein database PPDB (shown in comment field) and TargetP (shown in TargetP prediction field).
Figure 4 Data sources used in MetNetDB. MetNetDB collects and integrates biological network, protein annotation, gene annotation and metabolite annotation information from key online databases.
Figure 5. Visual pathway query displayed in MetNetDB curator tool window. A) An example of an input window when a curator has drawn a partial structure for a pathway query. B) This window displays a matching result from the structural query in panel A. The nodes of the input graph are designated by a red box. The dotted lines indicate the alignment between these nodes of the input graph and the pathway “ethylene biosynthesis and methionine cycle”. For instance, a dotted line indicates that S-Adomet node of the input graph is aligned to the S-Adomet in cytosol (light yellow background) of the pathway. In this pathway we can find a highlighted cycle that inexacty matches the input graph.
Figure 6 Data history browser. The data history browser window displays previous edit operations on biological network data. The curator can always retrieve, review and restore these history data from this window.
Figure 7. Tracking the history of changes is implemented to provide MetNetDB concurrency control during curation. MetNetDB uses version comparison to prevent multiple curators from editing different versions of the same data. At any instance, only one curator can update the database. Before an update can be made, the most current version must be the same as the one from which the curator initially retrieved the data, thus ensuring that the curator is updating the correct version of the data.
SUPPLEMENTAL DATA

The following materials are available in the online version of this article.

Table S1. Data sources of MetNetDB

Table S2. The subcellular location hierarchy defined in MetNetDB (Locations are based as possible on GO categories)

Table S3. The interaction types defined in MetNetDB

Table S4. Rules of the labeled graph model used in MetNetDB

Figure S1. Gene annotation from multiple sources

Figure S2. The annotation for the metabolite ethylene

Table S1. Data sources of MetNetDB

<table>
<thead>
<tr>
<th>Database</th>
<th>Format</th>
<th>Information retrieved</th>
</tr>
</thead>
<tbody>
<tr>
<td>AraCyc</td>
<td>Plain text files organized according to frame data model</td>
<td>Pathways, interactions and biomolecules participated in. Name, synonyms, references, comments. Majority metabolic pathways in MetNetDB come from AraCyc</td>
</tr>
<tr>
<td>AGRIS</td>
<td>Plain text files organized according to simple graph model</td>
<td>Transcription network, references and binding sites of individual transcriptional factors</td>
</tr>
<tr>
<td>GO</td>
<td>MySQL dump files organized according to acyclic directed graph data model</td>
<td>The whole copy of gene ontology database</td>
</tr>
<tr>
<td>TAIR</td>
<td>Plain text files (Tabular data)</td>
<td>Affymetrix array elements and their corresponding LocusID mapping, Unitprot ID, TargetP location of polypeptides, loci of each AraCyc pathway</td>
</tr>
<tr>
<td>MapMan</td>
<td>Excel files (Tabular data)</td>
<td>Gene annotation, MapMan BIN ID, gene function category</td>
</tr>
<tr>
<td>BioCyc open chemical compound database</td>
<td>Plain text files organized according to frame data model</td>
<td>UNIQUE-ID, synonyms</td>
</tr>
<tr>
<td>ChEBI</td>
<td>MySQL dump organized according to directed graph data model</td>
<td>ChEBI ID, formula, molecular weight, IUPAC, SMILES</td>
</tr>
<tr>
<td>PubChem</td>
<td>XML files organized according to object data</td>
<td>PubChem CID, synonyms</td>
</tr>
<tr>
<td>Database</td>
<td>Format</td>
<td>Information retrieved</td>
</tr>
<tr>
<td>-------------</td>
<td>---------------------------------------------</td>
<td>---------------------------------------------------------------------------</td>
</tr>
<tr>
<td>NCI</td>
<td>Structure data format according to object data model</td>
<td>Synonyms, CAS registry number</td>
</tr>
<tr>
<td>KEGG</td>
<td>Plain text files (for compounds) organized according to object data model</td>
<td>Synonyms</td>
</tr>
<tr>
<td>SUBA</td>
<td>Excel file</td>
<td>Protein subcellular location including experiment verified and software predicted</td>
</tr>
<tr>
<td>PPDB</td>
<td>Tabular data</td>
<td>Curated protein subcellular location, especially those in plastid</td>
</tr>
<tr>
<td>AMPDB</td>
<td>Tabular data</td>
<td>Mitochondrion proteins, the subcellular location comes from computational prediction</td>
</tr>
<tr>
<td>AtNoPDB</td>
<td>Tabular data</td>
<td>Nucleolar proteins, subcellular location comes from prediction and experiments</td>
</tr>
<tr>
<td>AraPerox</td>
<td>Plain text</td>
<td>Putative proteins in peroxisomes. Subcellular location comes from literature and computational prediction</td>
</tr>
<tr>
<td>plprot</td>
<td>Plain text files organized according to object data model</td>
<td>Subcellular location comes from TargetP prediction</td>
</tr>
<tr>
<td>BRENDA</td>
<td>Plain text files organized according to object data model</td>
<td>Enzyme’s interaction, substrate, product, activator, inhibitor, synonyms, metal ions, references</td>
</tr>
<tr>
<td>MetNet curator</td>
<td>Manually curation</td>
<td>All, with focus on signal transduction information</td>
</tr>
</tbody>
</table>
Table S2. The subcellular location hierarchy defined in MetNetDB (Locations are based as possible on GO categories)

<table>
<thead>
<tr>
<th>Subcellular location hierarchy</th>
</tr>
</thead>
<tbody>
<tr>
<td>plastid</td>
</tr>
<tr>
<td>plastid envelope</td>
</tr>
<tr>
<td>plastid outer envelope</td>
</tr>
<tr>
<td>integral to plastid outer envelope</td>
</tr>
<tr>
<td>peripheral to plastid outer envelope</td>
</tr>
<tr>
<td>peripheral to cytosolic side of plastid outer envelope</td>
</tr>
<tr>
<td>plastid intermembrane space</td>
</tr>
<tr>
<td>plastid inner envelope</td>
</tr>
<tr>
<td>integral to plastid inner envelope</td>
</tr>
<tr>
<td>peripheral to plastid inner envelope</td>
</tr>
<tr>
<td>peripheral to stromal side of plastid inner envelope</td>
</tr>
<tr>
<td>plastid stroma</td>
</tr>
<tr>
<td>plastid inner membrane</td>
</tr>
<tr>
<td>plastid outer membrane</td>
</tr>
<tr>
<td>thylakoid</td>
</tr>
<tr>
<td>thylakoid membrane</td>
</tr>
<tr>
<td>integral to plastid thylakoid membrane</td>
</tr>
<tr>
<td>peripheral to plastid thylakoid membrane</td>
</tr>
<tr>
<td>peripheral to stromal side of plastid thylakoid membrane</td>
</tr>
<tr>
<td>peripheral to luminal side of plastid thylakoid membrane</td>
</tr>
<tr>
<td>thylakoid lumen</td>
</tr>
<tr>
<td>thylakoid inner space</td>
</tr>
<tr>
<td>plastoglobules</td>
</tr>
<tr>
<td>plastid nucleoid</td>
</tr>
<tr>
<td>plastid ribosome</td>
</tr>
<tr>
<td>cytosol</td>
</tr>
<tr>
<td>mitochondrion</td>
</tr>
<tr>
<td>mitochondrial matrix</td>
</tr>
<tr>
<td>mitochondrial inner membrane</td>
</tr>
<tr>
<td>mitochondrial outer membrane</td>
</tr>
<tr>
<td>mitochondrial intermembrane space</td>
</tr>
<tr>
<td>mitochondrial crista</td>
</tr>
<tr>
<td>apoplast</td>
</tr>
<tr>
<td>autophagosome</td>
</tr>
<tr>
<td>endoplasmic reticulum</td>
</tr>
<tr>
<td>plasmodesmatal endoplasmic reticulum</td>
</tr>
<tr>
<td>rough endoplasmic reticulum</td>
</tr>
<tr>
<td>rough endoplasmic reticulum cisterna</td>
</tr>
<tr>
<td>rough endoplasmic reticulum lumen</td>
</tr>
<tr>
<td>smooth endoplasmic reticulum</td>
</tr>
<tr>
<td>smooth endoplasmic reticulum cisterna</td>
</tr>
<tr>
<td>smooth endoplasmic reticulum lumen</td>
</tr>
<tr>
<td>Golgi apparatus</td>
</tr>
<tr>
<td>GARP complex</td>
</tr>
<tr>
<td>Golgi lumen</td>
</tr>
<tr>
<td>Golgi membrane</td>
</tr>
</tbody>
</table>
Subcellular location hierarchy

Golgi stack
Golgi cis-face
Golgi trans face
Golgi vesicle
  ER-Golgi transport vesicle
  inter-Golgi transport vesicle
  trans-Golgi network transport vesicle
inner membrane
lipid particle
microbody
  microbody space
  microbody lumen
  microbody membrane
membrane
nucleus
  nuclear membrane
    nuclear inner membrane
    nuclear outer membrane
  nuclear lumen
  nucleolus
  nuclear body
  nucleoplasm
plasma membrane
vacuole
  vacuolar lumen
  vacuolar membrane
not assigned
### Table S3. The interaction types defined in MetNetDB

<table>
<thead>
<tr>
<th>Interaction types</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enzymatic reaction</td>
</tr>
<tr>
<td>Catalysis</td>
</tr>
<tr>
<td>Translation</td>
</tr>
<tr>
<td>Transcription</td>
</tr>
<tr>
<td>Composition-AND</td>
</tr>
<tr>
<td>Composition-OR</td>
</tr>
<tr>
<td>Diffusion</td>
</tr>
<tr>
<td>Transport</td>
</tr>
<tr>
<td>Channel-type facilitors</td>
</tr>
<tr>
<td>ATP-driven Transporters</td>
</tr>
<tr>
<td>PEP-dependent Transporters</td>
</tr>
<tr>
<td>Decarboxylation-driven Transporters</td>
</tr>
<tr>
<td>Electron-flow-driven Transporters</td>
</tr>
<tr>
<td>Light-driven Transporters</td>
</tr>
<tr>
<td>Mechanically-driven Transporters</td>
</tr>
<tr>
<td>Positive regulation (indirect or unknown mechanism)</td>
</tr>
<tr>
<td>Allosteric activation</td>
</tr>
<tr>
<td>Competitive activation</td>
</tr>
<tr>
<td>Covalent modification</td>
</tr>
<tr>
<td>Complex formation(yielding active protein)</td>
</tr>
<tr>
<td>Transcriptional activation (unknown mechanism)</td>
</tr>
<tr>
<td>direct coactivation</td>
</tr>
<tr>
<td>Translational activation</td>
</tr>
<tr>
<td>Indirect activation</td>
</tr>
<tr>
<td>Negative regulation (indirect or unknown mechanism)</td>
</tr>
<tr>
<td>Allosteric inhibition</td>
</tr>
<tr>
<td>Competitive inhibition</td>
</tr>
<tr>
<td>Covalent modification</td>
</tr>
<tr>
<td>Complex formation(yielding inactive protein)</td>
</tr>
<tr>
<td>Transcriptional inhibition (unknown) mechanism</td>
</tr>
<tr>
<td>direct corepression</td>
</tr>
<tr>
<td>Translational inhibition</td>
</tr>
<tr>
<td>Indirect inhibition</td>
</tr>
<tr>
<td>Degredation</td>
</tr>
<tr>
<td>Two-component regulators</td>
</tr>
<tr>
<td>Bind</td>
</tr>
<tr>
<td>Act as adaptor protein (specific case of binds)</td>
</tr>
<tr>
<td>Others (user submitted, curator evaluated)</td>
</tr>
</tbody>
</table>
**Table S4. Rules of the labeled graph model used in MetNetDB**

**Rules of the MetNetDB labeled graph model**
- A node must either be an biomolecule or an biochemical reaction
- Genes should be in nucleus, mitochondrion or plastid for plants
- An enzyme must be either a polypeptide or a protein complex
- The substrate and the product of a transportation must be the same biomolecule but in different subcellular locations
- A pathway graph should contain no duplicated nodes which represent same biomolecules in same subcellular location
- A pathway graph should contain no duplicated nodes which represents same biochemical reactions
- An interaction must have at least one substrate and one product
- There must be an interaction node between any two biomolecular nodes in a pathway graph
- No two interaction nodes can be adjacent except a catalysis node and an enzymatic reaction node
- A pathway graph should contain no orphan nodes which do not connect to any other nodes
- A biomolecular node should have a single subcellular location value
Figure S1. Gene annotation from multiple sources. For each locus and its gene product, the annotations from GO, TAIR, MapMan and AGRIS are displayed.
Figure S2. The annotation for the metabolite ethylene. The synonyms of ethylene come from KEGG, PubChem, AraCyc, ChEBI and NCI. The chemical formula, IUPAC ID, SMILES and external database cross-references are from ChEBI. MetNetDB curator tool is used for the visualization.
APPENDIX C. ATP-CITRATE LYASE TARGETED TO PLASTIDS OF ARABIDOPSIS THALIANA IS UNDER COMPLEX POST-TRANSCRIPTIONAL CONTROL

Heather L. Babka¹ and Eve S. Wurtele¹

ABSTRACT

Acetyl-coA is necessary for many biochemical reactions in living organisms, including the synthesis of fatty acids, sterols, and many secondary metabolites. Acetyl-CoA is synthesized within the compartments in which it will be utilized, as CoA moieties cannot cross membranes. ATP-citrate lyase (ACL) is responsible for synthesizing the cytosolic pool of acetyl-CoA. Native ACL has been targeted to the plastids of Arabidopsis thaliana using the transit peptide from the small subunit of Rubisco. While the introduced ACLA-1 and ACLB-2 genes are transcribed to a high level, this does not correspond to increased ACL protein or activity in planta. The results shown here indicate that ACL protein accumulation and activity are subject to post transcriptional control.

INTRODUCTION

Acetyl-CoA is a metabolite central to metabolism, juxtaposed between catabolism and anabolism. As biomolecules containing the CoA moiety cannot cross cellular membranes, acetyl-CoA must be synthesized within the compartment it is to be utilized. ATP-citrate lyase (ACL) is responsible for synthesizing the cytosolic pool of acetyl-CoA, via the ATP dependent cleavage of citrate into acetyl-CoA and oxaloacetate. The cytosolic pool of acetyl-CoA is utilized to elongate fatty acids and to produce a large number of specialized phytochemicals including flavonoids, sterols, and isoprenoids (Figure 1) [11, 25]. Until recently the plastidic pool of acetyl-CoA was thought to be provided by the combined action

¹ Department of Genetics, Development, and Cell Biology; Iowa State University
of the pyruvate dehydrogenase complex and acetyl-CoA synthetase [13, 17]. Much evidence now suggests that pyruvate dehydrogenase complex provides the bulk of plastidic acetyl-CoA, while acetyl-CoA synthetase (ACS) may be present as a bypass to prevent acetate toxicity or to shunt carbon into the glyoxylate cycle [18].

ACL has been found in yeast, fungi, human, Achaea, algae, bacteria, and a number of plant species including *Arabidopsis thaliana*, *Pisum sativum*, *Glycine max*, and *Brassica napus* [1, 2, 7, 9, 11, 12, 19, 22, 24, 27, 30]. In the *Arabidopsis* genome there are three genes that encode *ACLA* (*ACLA-1*, *ACLA-2*, and *ACLA-3*) and two that encode *ACLB* (*ACLB-1* and *ACLB-2*) [9, 10]. ACLA protein from *Arabidopsis* is similar to the N-terminus of rat and human ACL, while the ACLB protein is similar to the C-terminus of ACL protein from rat and human [8]. The ACL protein is localized to the cytosol in all species yet studied and is an octoheteromer, likely in an A₄B₄ configuration [9]. The RNA of *ACLA* and *ACLB* have been shown to coaccumulate with homomeric acetyl-CoA carboxylase. *Arabidopsis* plants expressing an anti-sense copy of *ACLA-1* exhibit a severe phenotype, indicating that there is only one source of cytosolic acetyl-CoA. Anti-sense *ACLA-1* plants are dwarf in size, dark green, have increased accumulation of both starch and anthocyanins, and show altered cuticular wax. This phenotype can be reversed by the application of exogenous malonic acid [10].

ACL can be targeted for expression in other compartments of *Arabidopsis* to boost the pool of acetyl-CoA. In particular, if plastids contain a pool of citrate, then increasing acetyl-CoA production via ACL could result in an increase of fatty acid biosynthesis. In *Brassica napus* developing seeds and *Glycine max* cotyledons the activity of ACL corresponds to the accumulation of fatty acids [23, 27]. *De novo* fatty acid biosynthesis is
initiated within the plastid of *Arabidopsis* via the carboxylation of acetyl-CoA into malonyl-CoA by the action of heteromeric acetyl-CoA carboxylase (htACCase). When rat liver ACL was targeted to the plastids of tobacco an increase in ACL protein content and activity was observed as well as an increase in fatty acid biosynthesis [26]. Acetyl-CoA production in *Arabidopsis* plastids could be boosted by targeted native ACLA and ACLB proteins to the plastid.

**MATERIALS AND METHODS**

**Vector construction**

The base vector used for plant transformation is pCB302-1 [31] *ACLA-1* was PCR amplified from pBS-ACLA from the Nikolau lab (Iowa State University, BBMB) with *Sma*I ends and ligated into pCB302-1, resulting in vector pCG-CB-ACLA. *ACLB-2* was removed from the vector pVBVYC01 using *Eco*RI, and ligated into pCB302-1, resulting in the vector pHJ210. Both of the resultant constructs were electroporated into *Agrobacterium* tumefaciens strain C58C1.

**Plant transformation and selection**

*Arabidopsis thaliana* glabrous Columbia ecotype (Co0-gll) (Lehle Seeds, Round Rock, TX) was transformed by dipping inflorescence stalks (approximately 40 days after planting) in infiltration medium containing *Agrobacterium tumefaciens* (strain C58C1) with either p35S:TP:anti-sense-ACLA-1 or p35S:TP:anti-sense-ACLB-2. Infiltration media contained 0.22% MS [20] salt mixture (Sigma, St. Louis, MO), 1X B5 vitamins, 5 % sucrose, 0.05 % MES-KOH, pH7.0, 44 nM benzylaminopurine, and 0.02 % Silwet L-77 (OSI Specialties, South Charleston, WV). Plants were placed at 22°C under continuous light (35 μmol m⁻² sec⁻¹) until seed harvest.
Transgenic plants containing p35S:TP:anti-sense-ACL-A1 were designated TPACLA and those containing p35S:TP:anti-sense-ACL-B-2 were designated TPACLB. Selection was performed by sowing seeds on sterile LC1 Sunshine Mix soil (Sun Gro Horticulture, Bellevue, WA) and selected by applying 0.5% glufosinate (Agr-Evo, Marysville, CA) 10 days later. Twenty days after sowing individual transformants were transferred to individual 2-inch pots. Plants were watered continuously with 21-nitrogen: 8-phosphorous: 18-potassium, fertilizer mix (“Nutriculture,” Plant Marvel Laboratories, Inc, Chicago Heights, IL).

**RNA extraction and Northern analysis**

Total RNA for Northern blots was extracted using a modification of the methods described by Kirk and Kirk [15]. Harvested plant tissue was ground in mortar and pestle using liquid nitrogen. Approximately 50 mg of tissue was placed in a microfuge tube and 1 mL of E-buffer was added (50 mM Tris-HCl, pH=8.0; 300 mM NaCl; 5 mM EDTA, pH = 8.0; 2 mM aurin tricarboxylic acid [ATA]; 2% SDS; 1% β-mercaptoethanol) along with 150 µL of 3 M KCl. The sample was first incubated on ice and then centrifuged to remove cellular debris and protein. Half a volume of 8 M LiCl was added to the supernatant and RNA was precipitated overnight at 4°C. The following day the RNA was centrifuged and resuspended in 0.1 mM ATA, and extracted once with phenol to remove any remaining protein. A final ethanol precipitation was carried out using two volumes of 100% ethanol and 0.1 volumes of 3 M NaOAc, and following centrifugation a pellet of pure RNA was recovered. The ATA protects the RNA from degradation, but because it irreversibly binds to the RNA any RNA purified in this manner is rendered useless for methods such as reverse transcription and microarray analysis.
RNA was denatured at 65°C and separated on a gel containing 1% agarose, 3% formaldehyde, and 10 mM phosphate, pH = 6.8 [6]. RNA was transferred to a positively charged nylon membrane (Hybond-N+, Amersham, Piscataway, NJ) using 20X SSC buffer [21]. Prehybridization and hybridization were carried out using a solution containing 7% SDS, 0.25 M sodium phosphate buffer, monobasic, 1 mM EDTA, and 1% casein [5]. The blots were hybridized with ^32P-labeled cDNA oligonucleotide probes corresponding to the full length *ACLA-1* or *ACLB-2* genes used for plant transformation. Radiolabel signal was detected using a PhosphorImager Typhoon 8600 (Molecular Dynamics, Sunnyvale, CA).

**Protein extraction**

Protein extraction was carried out on rosette leaf tissue homogenized in a liquid nitrogen cooled mortar and pestle containing extraction buffer (50 mM Tris-HCl, pH = 8.0; 1 mM EDTA, pH = 8.0; 10 mM dithiothreitol [DTT]; 1.5% [w/v] pre-swollen PVPP; 1 mM phenylmethylsulfonylfluoride, and 1 mM paraminobenzamidine). The resultant slurry was centrifuged at 4°C for 3 min at 12,000x g, and then for an additional 12-20 min at 4°C to further remove cellular debris. Bradford assay was used to determine protein concentration, and the protein was diluted with 5X Laemmle buffer, boiled to denature, and stored at -20°C until needed for Western analysis [3]. Protein extracts for ACL activity assays were desalted through Sephadex G-25-150 columns (200 μL extract on 1 mL Sephadex bed volume) with elution buffer (50 mM NaH₂PO₄, pH = 7.2; 1 mM MgCl₂; 0.1 mM EDTA, pH = 8.0; 1 mM DTT). Desalted extracts were diluted to 10% (v/v) using glycerol and stored in liquid nitrogen.
Western analysis

Proteins were separated using denaturing conditions through 12.5 % acrylamide [16], and transferred to nitrocellulose (Nitropure, Osmonics, Minnetonka, MN) using 25 mM glycine, 20 % methanol [29]. Gels were transferred using a voltage of up to 125 V for 175 volt-hours.

Western blots were incubated in TBST (10 mM Tris-HCl, pH = 8.0; 15 mM NaCl; 0.1 % Tween-20) + 3 % BSA at room temperature overnight. Anti-ACLA or Anti-ACLB serum [9] was added to the blots in new 3 % BSA supplemented TBST at a 1:500 dilution [14] and incubated 3-4 h at room temperature. Any unbound antibody was removed by three 10 min washes with TBST. The blots were then incubated with 2 X 10^5 cpm/mL 125I-Protein A (Amersham, Piscataway, NJ) in TBST supplemented with 3 % BSA for 2 h at room temperature. Three 10-min washes with TBST were used to remove any excess protein A. Radiolabel signal was detected using a PhosphorImager Typhoon 8600 (Molecular Dynamics, Sunnyvale, CA).

Spectrophotometric assay of ACL activity

ACL activity was determined by a coupled spectrophotometric assay [9, 28]. The ACL assay contained 100-150 μL of desalted protein extract (50 – 500 μg protein); 200 mM Tris, pH = 8.4; 20 mM MgCl2; 1 mM DTT; 10 mM ATP; 10 mM citrate; 6 units malate dehydrogenase; and 0.1 mM NADH. Background oxidation of NADH was determined by monitoring the decrease in absorbance at 340 nm in the absence of CoA, and the ACL reaction was initiated by the addition of 0.2 mM CoA.

This assay measures the oxidation of NADH during a reaction coupled to the ACL reaction: the conversion of oxaloacetate (a product of the ACL reaction) to malate by malate
dehydrogenase. ACL activity can then be calculated using the extinction coefficient of NADH (6.22 mM-1cm⁻¹). There is a 1:1 correspondence between the rate of NADH utilized and units of ACL activity.

**Plant crosses**

Pollen was removed from TPACLA overexpressing plants (T4 plants) and TPACLB overexpressing plants (T3 plants), 2 d after flowering (DAF). The pollen was placed on stigmas of TPACLB (0 DAF) overexpressing plants or on the stigmas of TPACLA (0 DAF) overexpressing plants, respectively. Siliques were grown to maturity; the F1 seed was collected, and analyzed for the overexpression of both *ACLA-1* and *ACLB-2* transgenes. Resultant plants were designated TPACLAB F1 progeny.

**RESULTS**

**Plastid targeted *ACLA-1* and *ACLB-2* RNA can be overexpressed in *Arabidopsis thaliana***

*Arabidopsis* transformed with pCG-CB-ACLA or pHJ210 were cultivated and analyzed for expression of *ACLA-1* and *ACLB-2* RNA respectively. Northern blot analysis confirms that *ACLA-1* and *ACLB-2* are overexpressed in plants that survived BAR selection (Figure 2). Plants overexpressing *ACLA-1* or *ACLB-2* targeted to the plastid do not show any visually noticeable phenotype.

Crosses were made between *Arabidopsis* plants overexpressing plastid targeted *ACLA-1* and *ACLB-2*. The resultant progeny were screened for the overexpression of both RNAs (Figure 3). Progeny that showed over expression of both *ACLA-1* and *ACLB-2* were further screened for ACL protein accumulation and activity.
Plastid targeted *ACLA-1* and *ACLB-2* RNA overexpressed in *Arabidopsis thaliana* does not correspond to increased levels of protein expression or ACL activity.

Arabidopsis plants overexpressing both *ACLA-1* and *ACLB-2* RNA were further analyzed at the protein level. As seen in Figure 4, neither ACLA-1 nor ACLB-2 protein accumulates to higher levels in transgenic plants overexpressing both *ACLA-1 + ACLB-2*. To further verify these results ACL activity assays were performed. ACL activity assays show that ACL activity in transgenic Arabidopsis hyper-accumulating *ACLA-1* and *ACLB-2* transcripts is not significantly different than wildtype ACL activity (p-value: 0.11) (Figure 5). The lack of ACL protein accumulation and the lack of ACL activity are consistent over many independent transgenic lines.

**DISCUSSION**

The 35S promoter has been successful at increasing the expression of both *ACLA-1* and *ACLB-2* RNA in *Arabidopsis thaliana*. The overexpression of one subunit (*ACLA-1* or *ACLB-2*) does not impact the expression of the other subunit. In previous studies, it had been shown that reducing the RNA accumulation of one subunit of ACL, *ACLA-1*, causes reduction in the RNA accumulation of the other ACL subunit RNAs [10]. When *ACLA-1* expression is reduced, there is also a decrease in ACLA protein and a proportionate decrease in ACLB protein [10]. In fact it is quite common to observe the subunits of multi-subunit enzymes being co-regulated. This has been observed with two subunits of MCCase, *MCC-A* and *MCC-B*, showing similar regulation at the mRNA and protein levels under various environmental conditions [4]. The subunits of heteromeric ACCase are also coordinately expressed at the mRNA level in *Arabidopsis* [14].
The increase in *ACLA-1* and *ACLB-2* transcript accumulation does not correspond to any change in the amount of either protein found in extracts. ACL activity in TPACLAB plant extracts does not show an increase in proportion to the increased RNA accumulation observed. Therefore, ACL must be regulated downstream of RNA processing, likely at the translational or post-translational level. This phenomenon should be further studied to better understand how post-transcriptional regulation can play a role in metabolic pathway regulation *in planta*. Rangasamy and Ratledge [26] successfully introduced rat liver ACL into tobacco plastids. Perhaps being a foreign gene, the human liver ACL was not subjected to the same regulation patterns as a native *Arabidopsis* gene.

Studies could be performed to determine if ribosomes are attaching onto introduced *Arabidopsis* *ACLA-1* or *ACLB-2* mRNA. If ribosomes are not successfully attaching to the introduced RNAs, then there is no chance to use this RNA overexpression method to boost acetyl-CoA production in plastids without altering the ability of ribosomes to bind to the introduced RNAs.

**REFERENCES**


Figure 1. Subcellular compartmentalization of acetyl-CoA metabolism. Reactions of acetyl-CoA metabolism in Arabidopsis. The *de novo* ACL reaction is shown in orange in the cytosol, while the other natural reactions are shown in black. The introduced ACL reaction in the plastid is shown in red.
Figure 2. Northern blot of Arabidopsis overexpressing ACLA-1 or ACLB-2 RNA chosen for crossing. Total RNA was extracted from Arabidopsis rosette leaves, run on a denaturing gel, transferred to a nylon membrane, and probed using cDNA for ACLA-1 or ACLB-2 labeled with $^{32}$P-dCTP. Panel A: Northern blot from plants containing plastid targeted ACLA-1 visualized using a Phosphoimager. Panel B: Northern blot from plants containing plastid targeted ACLB-2 visualized using a Phosphoimager. The black stars indicate plants over-expressing either ACLA-1 or ACLB-2. W indicates lanes containing wildtype RNA. Arrows indicate the location of the two subunits: ACLA-1 (A) and ACLB-2 (B).
Figure 3. RNA blots detecting \textit{ACLA-1} and \textit{ACLB-2} RNA in transgenic F3 lines containing plastid targeted ACLA and ACLB transgenes. Total RNA was extracted from \textit{Arabidopsis} rosette leaves, run on a denaturing gel, transferred to a nylon membrane, and probed using cDNA for \textit{ACLA-1} and \textit{ACLB-2} labeled with $^{32}$P-dCTP. Panels A and C: Northern blots from plants containing plastid targeted \textit{ACLA} and plastid targeted \textit{ACLB} transgenes, and visualized using a Phosphoimager. The black stars indicate plants over-expressing both subunits of ACL. W indicates lanes containing wildtype RNA. Panels B and D: rRNA visualized on the identical ethidium bromide stained, RNA gels prior to transfer. Arrows indicate the location of the two subunits: \textit{ACLA-1} (A) and \textit{ACLB-2} (B).
Figure 4.


**Figure 4. Western Blots detecting ACL-A1 and ACL-B2 polypeptides in transgenic F3 plants containing plastid targeted ACLA and ACLB transgenes.** Total protein was extracted from whole *Arabidopsis* bolts 35 DAP and 200 ug of protein was loaded, and run on 12.5% denaturing polyacrylamide gels. Protein was transferred to nitrocellulose. Membranes were then exposed to anti-ACLA or anti-ACLB antibody followed by exposure to the secondary antibody, \(^{125}\text{I}\)-Protein A. Panels A and D: Western blots of the ACLA polypeptide. Panels B and E: ACLB Western blots showing ACLB polypeptide. C and F. Coomassie Blue stained gel to show loading. The black stars indicate plants over-expressing RNA for both subunits of ACL in northern blots. (W) Protein from wildtype. Arrowheads indicate protein of interest, ACLA (A) or ACLB (B).
Figure 5. ACL activity of protein from wildtype and TPACLAB protein extracts determined via spectrophotometric assay. The assay measures the oxidation of NADH during a reaction coupled to the ACL reaction: the conversion of oxaloacetate (a product of the ACL reaction) to malate by malate dehydrogenase. There is a 1:1 correspondence between NADH utilized and units of ACL activity. p-value from a student’s T-test probability of a difference is 0.11.
ACKNOWLEDGEMENTS

First I would like to extend my thanks to Dr. Eve Syrkin Wurtele, my wonderful Ph.D. mentor. She has always challenged me to think in new ways and to try and see the big picture.

Dr. Harry Horner as been a wonderful mentor with respect to teaching and microscopy. I enjoyed our many discussions about research as well as all the times spent sitting with students on the SEM and STEM.

The crew of the Microscopy and NanoImaging Facility, Tracey Pepper and Randy DenAdel were quite helpful helping me run equipment and work out technical issues related to microscopy.

I am thankful for the technical and creative assistance Dr. Hilal Ilarslan provided for all of the TEM work.

Suh-Yeon Choi for all of her metabolomics expertise and for being the best officemate!

Dr. Matt Hillwig and Dr. Lankun Wu for all of their technical assistance working on the Hypericum paper. You guys also made my time in graduate school much more enjoyable.

I also thank Dr. Beth Winters-Rozema for the extremely helpful editing of this dissertation.

Lastly, my husband Scott for putting up with a stressed out graduate student and the wonderful care he gives Tatiana so I can finish this work. And of course Tatiana, my wonderful daughter, who has made dissertation writing an adventure.