Gene Expression in Plants: Use of System Identification for Control of Color

Terrence P. McGarty and Lloyd Nirenberg, *Member, IEEE*

***Abstract*—Plant color is controlled by the expression of certain chemicals in secondary pathways in plant metabolism, especially the anthocyanin pathways. The secondary pathways are controlled by protein concentration resulting from selective genes and their rate of expression. This paper analyzes these complex pathways as state machines and uses various system identification techniques to identify the mechanism of the pathways. Using the mechanism of the pathways provides a mechanism for control of the pathways to produce a desired color**

***Index Terms*—Gene Expression, Secondary Pathways, System Identification, Gene Control.**

I. Introduction

The genetic control of the expression of color in plant flowers is currently understood to be effected at several levels in the genetic and chemical pathways in plant cells. The most recent understanding is that there are secondary chemical pathways which produce such chemical products as anthocyanins and these pathways are controlled by enzymes, proteins, which are the products of genes in the plant cell. The plant flower color is then a result of the concentrations of the anthocyanins generated in these secondary pathways, for example. Each secondary pathway is rate controlled and is driven by enzyme concentrations. The greater the concentration of the enzyme, the greater the specific anthocyanin concentration. At the gene level other controlling factors as well. There are other secondary genes which may activate or suppress the gene which generates the activating enzyme. The combination of these elements can be expressed as a dynamic system and the process of determining the characteristics of this dynamic system can be posed as a classic system identification problem. That problem is analyzed in this paper and we focus on a specific species, Hemerocallis of the family Lilliaceae, a monocot plant found in China, Korea and Japan. The inverse of this problem is also considered, namely if we desire a specific color, then what do we seek to control at the gene level to effect this desired color. This is the controllability problem.

Manuscript received _____.

T.P. McGarty is with The Telmarc Group, Florham Park, NJ 07932 and with MIT, Cambridge, MA 02139 (tmcgarty@telmarc.com or mcgarty@mit.edu)

Lloyd Nirenberg is with The Telmarc Group, Florham Park, NJ 07932, (nirenberg@telmarc.com)

A. Overview

The following questions are addressed in this analysis and model development:

1. Given a dozen or more species plants which are relatively stable and consistent in the wild, how does the variation in color in hybrids arise. What is the cellular basis of color, and what is the genetic set of mechanisms which controls it?
2. Given the complexity of color, form and variegation in the hybrids, what is the genetic basis for the control mechanisms intracell and intercell? For example, how are such colorations as eyezones formed and what is the intercellular communications mechanisms which effect this?
3. Given what now appears to be a set of well-understood pathway- control mechanisms by enzymes produced within the cell and the gene control mechanisms for expression of these proteins, how are these combined to produce intra cellular coloration and what are the inter cellular communications which spread the colors out over the inflorescence?
4. Given that we can answer the above, can we generate a mathematical system for gene expression and control and using the model solve the coloration problem using system identification or inversion?
5. Given that we could solve the above problem, then how could we apply positive control to coloration and produce whatever color we desire?

Our approach in this paper is fairly straightforward. We focus on a specific genus, Hemerocalis, and on a specific part of the plant in that genus, the inflorescence (see [1]-[5] each of which provides an overview of this genus).

There are two areas which are developed herein. They are the characterization of flower color and the system structure of genetic control of secondary pathways.

Flower Color: We present an overview of the process of developing color in flowers. We present an overview of the

anthocyanins, flavonols, and carotenoids. We review their pathways and summarize recent research which had identified the enzymes on each link of the pathway and the genes controlling those enzymes. This has been accomplished over the past few years and is critical to the understanding of the overall system approach.

Cell Genetics: We provide a detailed overview of cell genetics and how activators and repressors are key elements in the overall expression of enzymes and in turn the development of color. We review the cell elements and especially the process of gene expression. We discuss activators and repressors and the mechanisms of their actions. Their existence results from the work of Monod and Jacob in the early 1960s (see [6]-[9] for an overview of the genetic control mechanisms for plants).

B. System Models for Gene Expression:

Recently the biological community has applied system models to biological systems. We build on that effort and develop modals for the expression of flower colors. We recognize that color is a result of a mixture of secondary plant products such as anthocyanins. We can from the color of a flower determine what the mixture of each anthocyanin is. The concentration of an anthocyanin is a result of the concentration of the enzymes in the pathway which produces the anthocyanin, and typically the lowest enzyme concentration is the dominant factor. We also know that the concentrations of the enzymes is a result of activators and repressors, proteins also generated in a cell, which turn on or turn off the enzyme controlling the pathway. The work in [10] provides an excellent review of the status of systems techniques currently employed in the genetic analysis of organisms.

Combining these ideas we can develop a top down system model for color. The output or observation equation is the color, and the system equation is a dynamic process wherein the states are the protein concentrations from a large enough set of gene expressions, wherein genes are allowed to control other genes via an nth order dynamic process. We also allow for uncertainty by adding a “noise” process which converts the overall system model into a linear dynamic stochastic system with observables. We extend that model from a single cell to a matrix of interconnected cells. This allows us to explore the processes one sees in the development of eyezones and other sharp transitions of color in flowers. We use models which have been previously studied for color variation and apply those to the flower. In particular we will focus on each of the biochemical elements shown below:

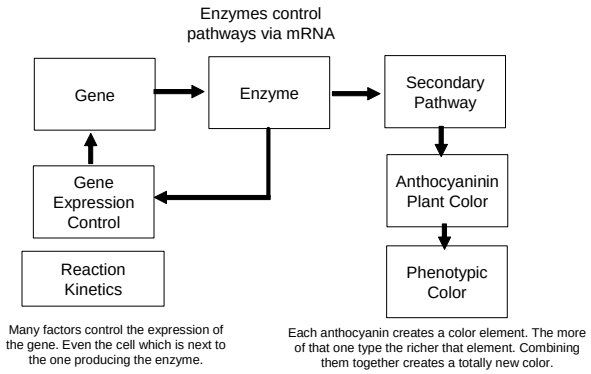

Figure 1: Basic Model for the production of secondary pathway agents controlled by underlying gene expression.

C. Prior Work

The key prior works fall into three categories; (i) underlying genetic studies and understandings of the genus, (ii) detailed elucidation of the control of pathways and the effecting gene sequences, and (iii) the development and application of models for the analysis and synthesis of gene expression.

1) Genetic Structure of Hemerocalis

Various recent works [11]-[13] have provided detailed genetic analyses of the genus. Specifically [12] has provided a detailed study of three populations of the species hakuensis and has shown that there is a significant intra-species variation. This has been known for many years. This was a problem for many plant sytematicists who had few examples of species available and used this limited number to describe the species. It is necessary to perform extensive field work to fully understand the intra-species variation.

Specifically [13] have studied the species citrina. Their work included a detailed analysis of certain exons and an understanding of the evolution of this species. They have begun to establish a basis for genetic analysis of within species characteristics. Extensive work in [14]-[16] have shown a detailed analysis of the full genetic variation in the genus using AFLP markers. They have also extended this to include many of the current common hybrids. Their key observation is that in the recent hybrids that they studied the genetic similarity has increased by approximately 10%. This demonstrates the rather interesting effect that if the genetic diversity is decreasing and the phenotypic changes are increasing then it clearly must be via expression.

2) Gene Expression and Pathway Control

Within the last five to ten years there has been considerable growth in the understanding of the control of the pathways which provide for color. The recent work in [17] provides a superb summary. The author reviews prior efforts and puts the

entire pathway management into perspective. [17] provides all controlling proteins and their causing genes. [17] does this for anthocyanins and flavones and isoflavonoids. The relationship to the abundance of effecting enzymes and anthocyanin expression as well as flavone expression is clearly demonstrated. This gives us a key in the development for our model. The work [18]-[19] predates [17] and is the seminal paper on genetic control of pathways. The authors in [18] have continued to publish their results on further clarification of the pathway management by gene expression.[20] was the first to publish the full pathway and that work is a major contribution to the field. Work on carotenoid pathways has been completed [21] and [22].

3) Modeling of Gene Expression: Analysis and Synthesis

The development of systems modeling for gene expression has most recently been exceptionally well articulated in [23]. In this work a collection of authors who are a mixture of systems experts and micro biologists present an up to date summary of all key works in this area. The work [24] is also an excellent modeling tool which applies a more Boolean approach to modeling the expression of genes. However the authors also extend their models to include mRNA and other pathways in a linear time varying system model as well. This latter approach coincides with the recent directions portrayed [23] and is consistent with the approach taken herein. [25] has provided a neural network approach to the understanding of gene expression. Although highly flexible this model is at best amenable for limited simulation analysis. [26] has provided a detailed systems model for expression using classic dynamic systems models. [27] has also provided a detailed dynamic model using their “differential equations” approach. We see that [26] and [27] have a great many similarities, as does the collection of authors in [23] but they all seem at best to be just becoming aware of the wealth of well understood theory in the control and estimation area [28].

II. Flower Color Expression

There is an extraordinary variation in the color of the hybridized flowers of the genus Hemerocallis. In a little over a hundred years hybridizers have taken the dozen or so species, all predominantly yellow, orange or red, intermixed them and as a result have created a very complex set of flowers with characteristics which differ dramatically from each of the species. The species have managed to maintain their separate identities over thousands of years but in a small fraction of time we have been able to introduce multiple forms and colors. To understand this process we first have to understand where the colors come from. How do we get purple from a plant which is red, yellow, orange and possibly even brown? How are the colors made? In this paper we focus on inflorescences of one color. The issue of variegated inflorescence has been studied initially in 1949 [29] in a brilliant paper before the Watson Crick model was developed and his analysis is expounded upon in Murray. The Turing model [29] is more complex than what we present here and will be detailed in another paper.

The first step in understanding that process is to understand the pathways that lead to color production in a single cell. Then we can address the issue of multiple cells and finally how the cells communicate. For example, how do we get an eyezone?. Why if a cell is white do we go so abruptly to a purple eyezone? What is the mechanism for this process? We begin the exploration of this issue with a analysis of the underlying pathways.

A. Pathways and Enzymes

Pathways are nothing more than a set of chemical reactions which get us from some primitive chemical to a more complex but useful chemical structure (see [60], [30], and [31]).In fact the pathways may be just a set of processes going from any one chemical structure to another independent of the nature of the starting and starting chemical. Some pathways are linear going from a beginning to and end and some are circular taking us from the beginning and back again (the Krebs cycle is an example). What makes the pathway work? Just three elements are required: (i) the underlying chemical constituents, (ii) some form of energy, (iii) generally some form of facilitation such as a catalyst and in our analyses this is an enzyme. We have the pathway but it is facilitated by an enzyme, a protein. The protein is generated by a gene. And the gene is activated by some other element, generally another protein. In our case shown below the output is some anthocyanin. The more of the enzyme, the more the gene expresses itself and the more anthocyanin we get; this is the basis of enzyme reaction kinetics [32]-[35]. Thus if we can get the gene to express then we get more of that specific anthocyanin, more pelargonidin for example. We defer to the next section how we get this gene to express so strongly. The opposite is also true: if we can suppress the gene then we can get less and even possibly no anthocyanin from the pathway. This is the first step in the development of an overall system model.

B. Anthocyanins

Let us consider our first pathway. This is the pathway which creates anthocyanins (see [17], [19], [20], and [36]). The anthocyanin molecule is shown below. Note on the B ring we have six sites to which we can attach differing molecular chains. This will be an important element when we see the different configurations and their implications. The anthocyanin or anthocyanidin molecule comes from two different pathways (see [37]-[39]). One is from the shikimic pathway and the other from the malonate pathway. This means that we have to understand both pathways to understand the ultimate abundance of the product.

In the Table 1 below we list the anthocyanidin and its resulting color. Each is obviously named after their related flower and each results an anthocyanin of a different color.

TABLE 1: ANTHOCYANIN AND ASSOCIATED COLOR RESULTING FROM ITS ACTIVATION

Anthocyanidin	*Color*
Pelargonadin	orange-red
Cyanidin	purplish red
Delphinidin	bluish purple
Peonidin	rosy red
Petunidin	purple

Each of the colors is the weighting of a red, green and blue combination which best matches the color. Thus one can, in an 8 bit color scheme for example, as one would find in any PC color scheme, get the resulting anthocyanin colors by blending the R, B, G elements to yield what we seek. This relating the colors back to RBG is critical since it get reflected in the ultimate flower color.

Now if we assume we have only anthocyanins for color, and that we have the above combinations available, we ask how do we combine these colors in a weighted manner to obtain the desired color. This approach is critical to our overall understanding. First we show by a weighted RBG we get the color we seek or the color which is presented. Then we assume that if we can then do the same for each anthocyanin, then we can create any desired color from a weighted collection of anthocyanins. This means that we can determine what the relative percents of expression of any anthocyanin are and this lets us go back to how strongly the gene for that anthocyanin is expressed. The model we presented earlier will be a key element in this overall process.

Now let us start with a simple expression (see [40], [41], and [42]). This text presents a detailed analysis of how color is characterized in the red/green/blue model. This model carries over directly to the computer color model that is currently in use. We have used it as a core baseline for the observable from the overall secondary pathway process. Thus for any color we can write:

$$\text{Color}=w_1\langle\text{Red}\rangle+w_2\langle\text{Blue}\rangle+w_3\langle\text{Green}\rangle$$

$$\begin{aligned}\langle\text{Red}\rangle &= \text{color base of Red}\\ \langle\text{Green}\rangle &= \text{color base of Green}\\ \langle\text{Blue}\rangle &= \text{color base of Blue}\end{aligned} \quad (1)$$

and if we define:

$$w=\begin{bmatrix} w_1\\ w_2\\ w_3\end{bmatrix} \quad (2)$$

$$s=\begin{bmatrix}\langle\text{Red}\rangle\\ \langle\text{Green}\rangle\\ \langle\text{Blue}\rangle\end{bmatrix}$$

resulting in:

$$Color = w^T s \quad (3)$$

Likewise we could state this by means of some combination of anthocyanins and their related colors. Now we can define any color as a combination of the three anthocyanin concentrations and the concentrations effects on color by using the basic red, green and blue bases as below:

$$Color=m_1\langle Pelargonidin\rangle+m_2\langle Delphinidin\rangle+m_3\langle Cyanidin\rangle$$

$$\begin{aligned}\langle Pelargonidin\rangle&=g_{1,P}\langle Red\rangle+g_{1,P}\langle Blue\rangle+g_{1,P}\langle Green\rangle\\ \langle Delphinidin\rangle&=g_{2,D}\langle Red\rangle+g_{2,D}\langle Blue\rangle+g_{2,D}\langle Green\rangle\\ \langle Cyanidin\rangle&=g_{3,C}\langle Red\rangle+g_{3,C}\langle Blue\rangle+g_{3,C}\langle Green\rangle\end{aligned} \quad (4)$$

We can measure the coefficients g in the above using standard colorimetry. If we define a matrix G as follows:

$$G=\begin{bmatrix} g_{1,P} & g_{2,P} & g_{3,P}\\ g_{1,D} & g_{2,D} & g_{3,D}\\ g_{1,C} & g_{2,C} & g_{3,C}\end{bmatrix} \quad (5)$$

Then we can determine the color as a simple product and the m values can be determined directly.

$$\begin{aligned}Color&=m^T Gs=w^T s\\ m&=G^{-1}w\end{aligned} \quad (6)$$

The above analysis shows us that we can analytically determine the expression of the anthocyanins from the color of the cell by means of the above formulas. Color is determined in the Red, Blue, Green approach by weighting each of this prime colors by some weight w. Also we can obtain the same color by weighting the alternative colors as associated with the anthocyanins present by a similar weight in this case m and m is directly related to the concentration of that anthocyanin. These are relative expressions but by benchmarking any one element we can make them all absolute in the cell as well.

C. Other Color Elements

Anthocyanins are not the only elements which are secondary products which produce color. There are three classes of chemicals which give rise to color; anthocyanins, flavones or flavonols, and carotenoids. The Table 2 below depicts the different elements and their colors. The approach we took above for the anthocyanins can be used for the flavones and carotenoids as well. It should be noted that there may not be a unique solution here but there are several possible. The solutions can be narrowed down by actual determination of one to three elements as baseline. The other two general classes are the carotenoids, the orange colors, and the flavonoids, the more white type of colors. We summarize the colors in the following Table 2.

TABLE 2: SUMMARY OF THE THREE MAJOR CLASSES AND THEIR AGENTS AND RESULTING COLORS.

Class	*Agent*	*Color*
Anthocyanidin		
	Pelargonidin	orange-red
	Cyanidin	purplish-red
	Delphinidin	bluish-purple
	Peonidin	rosy red
	Petunidin	purple
	Malvinidin	
Flavonol		
	Kaempferol	ivory cream
	Quercetin	cream
	Myricetin	cream
	Isorhamnetin	
	Larycitrin	
	Syringetin	
	Luteolin	yellowish
	Agipenin	Cream
Carotenoids		
	Carotene	orange
	Lycopene	Orange-red

D. Pathways

In this section we present the pathways for the three classes we have described above. We first present an overview of the pathway and then we present the details of the pathway and the enzymes used in each step. The key observation is that we must have enzymes to go from step to step in the pathways and that if any one enzyme is missing we cannot proceed on that path, and further the path with the small amount of enzyme becomes the limiting path. Thus, we do not have a one to one map here. The production of any one anthocyanin, for example, if limited by the lowest produced enzyme, and the other enzymes may be present in abundance.

The following is the overall pathway for all elements.

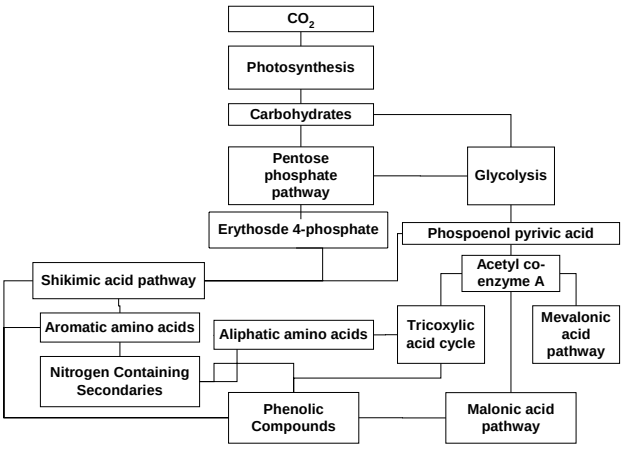

Figure 2: Pathway for the production of anthocyanin from Carbon Dioxide.

The above shows how we start from CO_2 and then go through a variety of other pathways. We will review those pathways in some detail since it is the enzyme control in them which is key. The pathway moves forward at a reaction rate which is determined by the concentration of the reaction supporting enzymes. This is modeled by standard enzyme reaction kinetics as is provided in [32], [43], and [10].

1) Anthocyanin Pathway

The anthocyanin pathway with the controlling enzymes is shown in Figure 4. The enzymes are presented in the arrows linking each step in this pathway. This pathway shows the start as a sugar element and then goes to phenyanaline and then down through the chain to one of the four indicated anthocyanins.

FIGURE 4: SPECIFIC ANTHOCYANIN PATHWAY FOR SPECIFIC PRODUCTS.

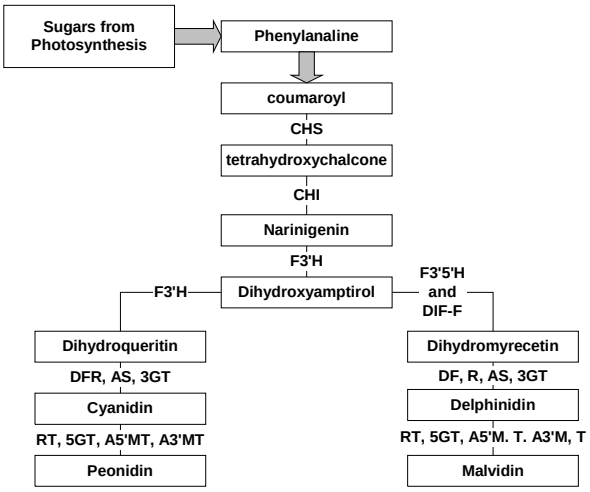

Figure 3: Specific Pathways for specific anthocyanins and the gene control via enzymes in specific pathway transitions.

Note that at each step there is an enzyme element. The genetic loci for cloned flavonoid enzymes in Arabidopsis are shown in the following Table 3

TABLE 3: ACTIVATING ENZYMES AND THE ASSOCIATED CHROMOSOME AND GENE LOCATION

Enzyme	Locus	Chromosome	Map Position
CHS	tt4	5	7,050 kb (MAC12)
CHI	tt5	3	21,000 kb (T15C9)
F3H	tt6	3	19,600 kb (F24M12)
F39H	tt7	5	4,400 kb (F13G24)
FLS	fls1<Enc	5	FLS1: 4,700 kb (MAH20) FLS2-5;: 32,150 kb (MBK5) FLS6: 24,350 kb (MRH10)
DFR	tt3	5	23,800 kb (MJB21)
LDOX	tt19	4	16,900 kb (F7H19)
LCR	ban,ast d	1	26,800 kb (T13M11)

What these process point out can be summarized as follows:

1. There are common pathways which are operational in all plants for the generation of the pigments.

2. Enzymes used as activators modulate the amount of production of the enzymes.

3. The products of these pathways, the anthocyanins, are driven by the concentration of the facilitating enzymes.

Secondary products always have this type of production process. As we look at a cell, from a system point of view we see facilitating proteins and secondary products. The concentration of the secondaries are proportional, in some general way, to the concentration of the facilitating proteins. However we see there are many facilitating proteins which may make this a more complex analysis, however doable.

2) Carotenoid Pathway

The carotenoid pathway is shown in [30]. It is similar in many ways to the anthocyanin.

3) Flavonol Pathway

The flavonol pathway is identical to that of the anthocyanin and is detailed in the work of [17].

III. Expression Analysis and Implications

In this section we develop a systems approach to the problem of color analysis and synthesis. This work is based upon the recent work [23] and also builds upon the work in McGarty (1971) which focused a systems approach to the overall identification problem.

A. Approach: Engineering versus Science

The approach we take in this paper is an engineering approach rather than a biological approach (see [24]-[26]). Our interest is in developing a model or sets of models which allow us, by a verifiable means, to show how the genes react and interact to produce the plant colors. We can compare this to the engineering approach to circuit design of transistor circuits versus the science of understanding the semiconductor from the point of view of detailed quantum mechanical models. The biologist in our approach is akin to the physicists. Our approach is akin to the engineer knowing that there is some set of physical processes inside the semiconductor which may clearly be important to the physicist but the engineer's interest is in designing and analyzing the circuit element to produce a system that behaves in a specific way. Thus for an engineer, if we increase a current here we get a decrease or an increase at some other point. The engineer creates a world view of a macro set of processes and models the details of the biologists in our case with a few set of equations which show the results of increases and decreases. One must be able to make measurements to show that the processes predicted indeed occur, to a reasonable degree of accuracy. Then one can analyze a genetic circuit and then in addition one can design a genetic circuit. We then can understand where the colors come from and possibly engineer the genes to develop and deliver on colors we desire.

B. A Control Paradigm

The expression regulator for any gene may be an activator or suppressor gene. It may be a result of a gene expression in the cell itself or quite possibly as we shall discuss fed through from another cell. There are many of these regulatory cycles and they are all interconnected. This basic paradigm is one of hundreds or thousands of such interconnected flows.

In developing our models we will use this construct. However, we can frequently focus on natural clusters of related genes. They may be a dozen or more such related genes in each cluster and possibly hundred of such clusters. Although cells and their proteins may affect all other cells, only a few of the genes regulated have a significant level of regulation. The low levels of "regulation" we shall consider just as noise.

C. A Model for Secondary Production

A system model for the relationship between the genes, proteins and secondary path chemicals can be developed. We assume we know or can determine the following:

1. The secondary pathway chemical steps are known. This includes what enzymes modulate what transitions in the pathway.
2. The resulting concentrations of the products from the secondary pathways are proportional to the concentrations of the enzymes acting as catalysts on the pathways. Namely the pathways follow known enzyme reaction kinetics.
3. The concentrations that result from the secondary pathways are reflective in the phenotype characteristic perceived.
4. That the phenotype characteristics perceived are measurable and can be analytically related to concentrations in secondary pathway products.
5. The genes which produce enzymes which modulate secondary pathways are known in detail.
6. That activator and repressor genes of the modulator genes are known or knowable. Their specific effect on the modulating gene does not have to be known a priori.
7. There exists a database of genes which are modulator, repressor, and activator characteristic from which one may be able to analyze their levels of expression using microarray or similar technologies.

In the event that these assumptions are valid, which is the case now for many plants, as well as many animal models including humans, then one can develop models to determine the "system model" of the genes and the secondary pathway elements. This is the "analysis" portion of the system. It provides the elements of the system and quantitative values for its dynamic behavior. The second portion is the "synthesis" portion. In the synthesis portion we assume we have determined the values to model the dynamics of the system. Next we seek to drive the system to a desired state, in this case a desired flower color. To many this is the "blue rose" or the "blue daylily" issue. We commence with a model for the gene and its control, the secondary pathway and complete the analysis issue. The system model is depicted in the following Figure 4.

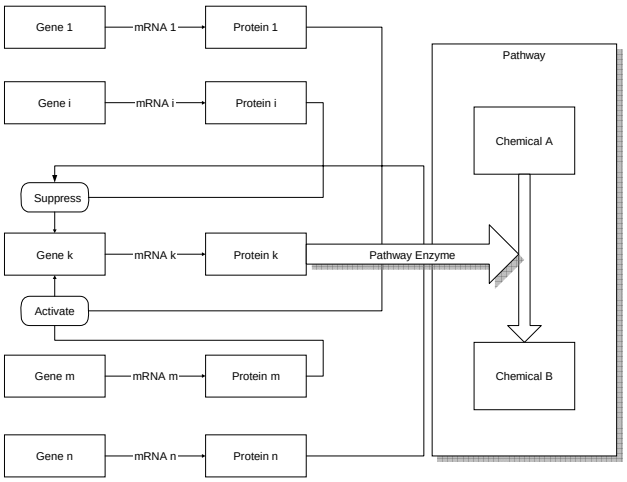

Figure 4: Model of pathway control and linkages between activating and suppressing genes.

The Equation (7) is an example of a general form for the expression of genes by means of the concentrations of the proteins related to specific genes. In the equation the expression of n genes is given as an n X 1 vector and the time variation of that vector is expressed as the sum of three other vectors; one is a vector which is dependent in a nonlinear manner on the concentrations themselves as well as time, one which is a function of some external set of influences which is a q x 1 vector u as well as time and an n x 1 vector which acts like random noise which accounts for a combination of noise and uncertainties.

$$\frac{dx}{dt} = f(x,t) + g(u,t) + n(t) \qquad (7)$$

$$x_k = \left[\text{Concentration } P_k\right]$$

A system model for the secondary pathways may also be developed. In this embodiment the following equation is an example of a general form for the secondary pathway. The m x 1 vector z represents what is observed and what in turn produces the phenotype characteristics. In the case of flower color the z elements may be the anthocyanin concentrations for example. In the equation below the z vector is a nonlinear function of the protein or enzyme concentrations plus added random noise which represents uncertainties as well as natural external disturbing measurement phenomenon, where we have used w(t) a white noise process to account for both measurement model inaccuracies and measurement errors.

$$z(t) = h(x(t),t) + w(t) \qquad (8)$$

The general models used above in this embodiment may be simplified by using standard techniques of linearizing them. The result of such a standard linearization process is

shown in the following equation. In the equation below the x vector is the protein concentrations resulting from genetic pathways and the z vector is the concentrations of the anthocyanins or in general they are the concentrations resulting from the secondary path chemical products. In this model the relationship of the protein concentrations to secondary chemical concentrations is assumed known or knowable. In this embodiment the elements depicting the dynamics of the protein concentrations are assumed unknown but can be ascertained quantitatively by means of the methods discussed herein.

$$\frac{dx}{dt}=Ax+\sum_{i=1}^{N} g_i p^T D_j x+o(x)+g(u)+v$$

$$z=h(x)+w=Cx+w \qquad (9)$$

$$z=\begin{bmatrix}[\text{Pelargonidin}]\\ [\text{Delphinidin}]\\ [\text{Cyanidin}]\end{bmatrix}$$

The above model may be further linearized to yield a simple linear system model which can be used in this embodiment. This simple linear model is shown in the following set of equations. The vector x is the set of protein concentrations resulting from the set of n gene expression interactions. The matrix a depicts the interaction between all of the genes as activators or repressors. the vector u is a known or unknown independent driving vector to the gene expression product. In this embodiments the matrix A will be determined by means of the procedures provided in this embodiment. The second equation depicts the steady state solution of the first equation. It should be noted that the steady state model is acceptable for plant color but the dynamic model may be required in many other systems where there is a dynamic portion to the system such as when in a human hormone release is involved. For the most part, however, steady state is adequate.

The steady state solution depicted below use the vector u and the inverse of the matrix A. In the equation below the vector x is composed of protein concentrations which effect secondary pathways such as the ones for the anthocyanins as well as genes which activate or repress the genes which directly express for color as shown in the equation below. In the equation below the concentration of secondary chemical products, z, are shown to relate to the concentrations of the proteins resulting from the first process. The relationship is via a matrix C which is in the equation below. Finally in the equations below the weight elements of the color expression, the vector elements denoted by m, are shown to be obtained from the concentrations of the secondary products.

$$\frac{dx(t)}{dt}=Ax(t)+u(t) \qquad (10)$$

For the steady state solution we have:

$$x=A^{-1}u \qquad (11)$$

$$\begin{bmatrix}x_1\\ x_2\\ x_3\\ x_4\\ x_5\\ x_6\\ \dots\end{bmatrix}=\begin{bmatrix}\textit{Concentration Enzyme effecting Pelargonidin}\\ \textit{Concentration Enzyme effecting Delphinidin}\\ \textit{Concentration Enzyme effecting Cyanidin}\\ \textit{Concentration Activator of Pelargonidin gene}\\ \textit{Concentration Repressor of Pelardonidin gene}\\ \textit{Concentration Activator of Delphinidin gene}\\ \dots\end{bmatrix}$$

The measurements can be described as follows:

$$z=Cx=CA^{-1}u \qquad (12)$$

$$\begin{bmatrix}z_1\\ z_2\\ z_3\end{bmatrix}=\begin{bmatrix}\textit{Concentration of Pelargonidin}\\ \textit{Concentration of Delphinidin}\\ \textit{Concentration of Cyanidin}\end{bmatrix}$$

From this we can obtain the relationships:

$$m_1=\kappa_1 z_1$$
$$m_2=\kappa_2 z_2$$
$$m_3=\kappa_3 z_3 \qquad (13)$$
$$or$$
$$m = Kz, and\ z=K^{-1}m$$

In the above we have assumed the following:

1. The system is at steady state. The expansion to a dynamic system is possible but it is unrealistic for plant colors. It functions in dynamic processes such as blood chemistry and endocrinology.
2. There is an unknown matrix A which we ultimately desire to obtain based upon the measurements made. This is the basis of the system identification problem.
3. There is a constant vector u which is the driver for the system.
4. There is a known relationship between the color elements in the space of colors using anthocyanin elements which can be used to determine the concentration of the anthocyanins. This is the K matrix and we assume that it is a diagonal with known values.

   The experimental data approach we use for the system identification process is the microarray. We assume we have a large collection of genes which have been

sequenced for the targetd flower. We also assume that we can create a microarray for these know genes and then using the array take samples of many different colors and test them in the array for the gene expressions. We also assume that we can determine expression intensity by measuring the expression intensity in each microarry cell (see [44] and [45]).

In the following Figure we depict a microarry which is composed of cells, one in each row-column pair. Each cell contains a row cDNA sample from a specific phenotype, and a column sample from a specific color, C. The cell then can be measured as to the intensity of the expression of the specific cDNA for each color. These measurements then become the basis of the data set.

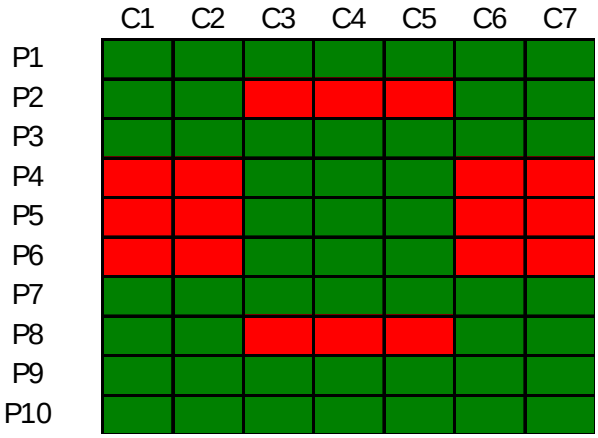

Figure 5: Example of a Microarray output showing activated cells for specific gene and color pairs. The rows are gene elements and the columns are form specific plant flowers.

By using standard means of microarry analysis (see [44], [45], [10], and [39]) as discussed above, the concentration of the protein may be determined. Simply, by having the row represent a know gene from the plant flower, and the columns being for sample from each color, we can measure the relative concentration of the products of genes for each color. This will be a key element in our determination of the system control parameters.

This concentration for the cell is denoted by the variable x. The phenotype color element is processed in a spectrum splitter which uses standard technology to determine the matrix elements of the Red, Green, and Blue elements. Using a color inversion matrix as described previously as G in this embodiment the vector weights for the secondary pathway chemical concentrations are obtained as denotes as vector elements m. Using a concentration matrix inversion processor the concentrations of the separate secondary chemicals denoted by vector elements c, in the case of Figure below.

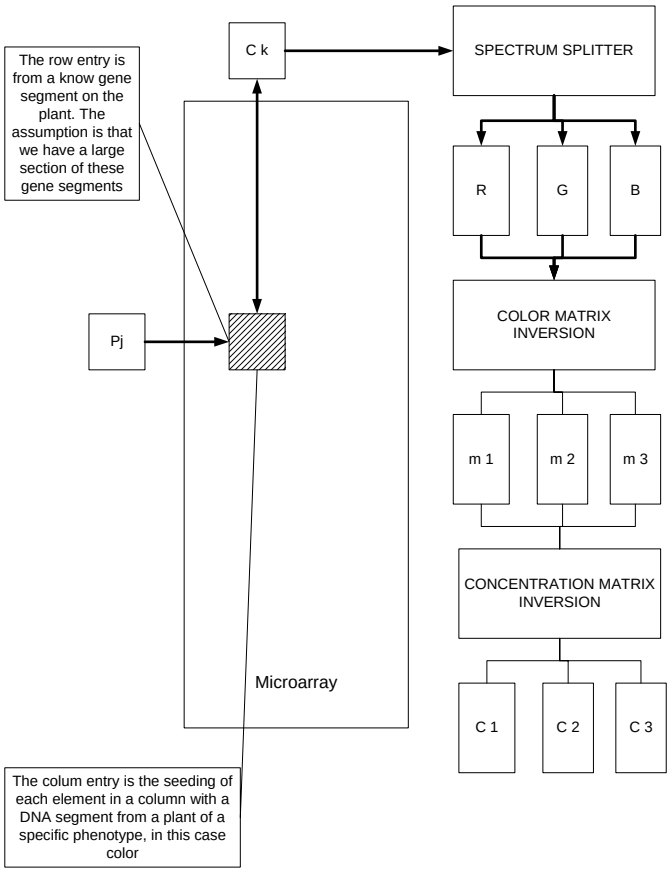

Figure 6: The models for the control of secondary products and their resulting phenotype results and the underlying gene control through enzyme management of the secondary pathway.

Specifically we perform the following steps:

1. Collect samples from various phenotypes, in this case color patches from the flower.
2. Prepare the cDNA from the plant
3. Prepare a microarray using the cDNA versus the various phenotypic color elements.
4. Run microarray analysis to obtain expression for each cDNA and phenotype cell entry.
5. Calculate the gene expression values for each microarray entry and denote the expression as a concentration element x(i,k), which is a measure of the concentration of protein element in row k for each color sample i.
6. Calculate color components for each color entry column and denote these in terms of standard R, G, B elements.
7. Calculate secondary pathway concentrations of their products, such as anthocyanins, based on the determined R, G, B elements and denote these secondary pathway product concentrations as z(i,k).
8. The above steps provide the basis for the analysis procedure to determine the system control constants, namely the system identification problem.

D. System Identification

The next step is to use the data obtained to estimate the constants we have assumed exist in the model for the modulator, repressor, activator genes and the control of the secondary pathway.

A set of measurements from the microarray data are collected and are presented in the equations below. The measurements consist of a vector composed of two sets of data; the phenotype elements and the gene expression concentration elements as described above. The equations below depict the measurements consisting of a collection of phenotype elements, m, and gene expression concentrations, x, for the entire collected data set. Namely we have a measurement tuple for every color column denoted by:

$$\{m_1, m_2, m_3, x_1, ..., x_n\} \tag{14}$$

The notation depicting this data set is further described in the equations below. The model employed in this embodiment then can show that the concentrations are related to the unknown expression matrix entries as is shown in the equation below relating x to the elements a of A and the elements of u. The elements a of A are defined in the equations below. Using standard matrix inversion methods the inverse elements are defined in the equations below and this permits the expression of the phenotypic secondary elements in terms of the matrix inverse elements as is shown in the equations below.

$$\begin{aligned} &\{m_1, m_2, m_3, x_1, ..., x_n\} \\ &or \\ &\left\{m^k{}_1, m^k{}_2, m^3_k, x^k{}_1, ..., x^k{}_n\right\} \quad k = 1.....N_{measure} \end{aligned} \tag{15}$$

Using the steady state model we obtain for the x values of concentrations the following:

$$x_j = \sum_{i=1}^{N} a^{ji} u_i \tag{16}$$

where we have defined the inverse by the terms shown as follows:

$$A^{-1} = \begin{bmatrix} a^{11}....a^{1n} \\ ..... \\ a^{n1}....a^{nn} \end{bmatrix} \tag{17}$$

$\{m_1, m_2, m_3, x_1, ..., x_n\}$

or

$\left\{m^k{}_1, m^k{}_2, m^3_k, x^k{}_1, ..., x^k{}_n\right\} \quad k = 1.....N_{measure}$

$$x_j = \sum_{i=1}^{N} a^{ji} u_i$$

$$A^{-1} = \begin{bmatrix} a^{11}....a^{1n} \\ ..... \\ a^{n1}....a^{nn} \end{bmatrix}$$

$$m_n(nm) = \sum_{i=1}^{N} \sum_{j=1}^{N} \kappa_{ni} a^{ij} u_j : n = 1,3$$

The next step is the calculation of the unknown matrix A from the collected data set of m values and x values specified. The first step in the process is the definition of the unknown elements of the matrix A as a vector of n^2 elements. The method chosen is a least squares fit method using a sequential procedure for obtaining the optimal fit from the data from each of the microcell elements. In the equations below the unknown elements are depicted as an n^2 vector. The objective is to determine an estimate of each entry and the vector estimate is denoted by $\hat{a}$. The least square means is one which minimizes the squared difference between the actual measured data as obtained and what the estimator predicts the data element should be using the most recent best estimate of the values a. This is stated mathematically in the equations below. Using standard mathematical techniques, namely a Newton method for solving the optimality problem, this is shown below as being the equivalent to solving a set of equations of a variable p (see [46] p. 111 for description using the Newton method as is done here). We thus seek to find an a vector as below:

$$\hat{a} = \begin{bmatrix} \hat{a}_{11} \\ .. \\ \hat{a}_{1n} \\ .. \\ \hat{a}_{n1} \\ .. \\ \hat{a}_{nn} \end{bmatrix} = \begin{bmatrix} a_1 \\ . \\ . \\ . \\ . \\ . \\ a_N \end{bmatrix} \tag{18}$$

to minimize the following metric:

$$min\left[\sum_{i=1}^{M} (\hat{m}_i - m_i)^2 + \sum_{j=1}^{N} \left(\hat{x}_j - x_j\right)^2\right] \tag{19}$$

where we define an h function as:

$$h(a)=\left[\sum_{i=1}^{M}(\hat{m}_i\text{-}m_i)^2+\sum_{j=1}^{N}\left(\hat{x}_j\text{-}x_j\right)^2\right] \quad (20)$$

and we seek a stationary point as follows:

$$\frac{\partial h(a)}{\partial a_n}=p_n(a)=0, n=1...N \quad (21)$$

where the stationary point is defined as:

$$p(a)=\begin{bmatrix} p_1 \\ .. \\ p_N \end{bmatrix}=0 \quad (22)$$

Using standard mathematical procedures, the equations in the previous portion of the embodiment may be solved in the following set of equations. The objective is to find the values of a which make the vector p zero. To accomplish this in the following equations a matrix K is obtained by the mathematical procedure depicted in the equations. The estimate of a after k+1 sampled from a microarry depicted in Fig 4 is shown as a function of the estimates after k samples from the same array. This procedure is performed sequentially for all samples in the array as shown in Fig. 4.

We desire to determine the a to solve:

$$p(a)=0 \quad (23)$$

We can define a matrix as follows:

$$K(a)=-\left[\frac{\partial p(a)}{\partial a}\right]^{-1}$$
$$\textit{where we define}: \quad (24)$$
$$\left[\frac{\partial p(a)}{\partial a}\right]=\begin{bmatrix} \frac{\partial p_1}{\partial a_1}.....\frac{\partial p_1}{\partial a_n} \\ ..... \\ \frac{\partial p_n}{\partial a_1}.....\frac{\partial p_n}{\partial a_n} \end{bmatrix}$$

with the solution can be sequentially determined as follows:

$$\hat{a}(k+1)=\hat{a}(k)+K(\hat{a}(k))p(\hat{a}(k)) \quad (25)$$

The process for solving the optimization is depicted in the following set of equations. The process commences with an initial guess for the vector a. Then the iterative process begins. A special method of elimination is proposed between steps. The method of elimination is one where genes which are not expressed may be eliminated from the sample. The determination of a non-expression gene can be determined from the process of stating that a gene is non expressing if it does not express or expresses below a set threshold level in any cell. We calculate the K gradient as shown in the equations below. We use the K matrix and calculate the p vector for the next steps as is shown in the equations below. We iteratively use the results of one step to determined the next best estimate. This process proceeds until complete. When complete, the best estimate of the elements of A is determined. The matrix K and the vector p are all evaluated by means of the difference values shown in the following equations (see [46] for examples of such optimization and see [47] with the same approach except using an estimation optimization. Also system identification techniques are the same as in [48] and [49]).

$$\hat{a}(0)=a^0 \quad (26)$$

which is the initial guess. The we use the first data tuple which we obtain from the microarry data as it may be normalized:

$$\hat{a}(1)=\hat{a}(0)+A(\hat{a}(0))g((\hat{a}(0)) \quad (27)$$

The difference elements are determined as follows and used:

$$\hat{x}_k(0)\text{-}x_{k,measured}(0)$$
$$\text{and} \quad (28)$$
$$\hat{m}_k(0)\text{-}m_{k,measured}(0)$$

as the data entry element for each of the data elements and where the estimates are calculated using the data collected from the system equations.

E. Modification for on/off A/R Genes

There is a slight modification we must include to deal with genes being on or off. We must return to the beginning to best understand it. Namely if the a's in the system matrix are all constant then by definition the colors remain the same. However, if any one or more of the A/R genes are on or off then we can get variation. We first explore the implications of this and then we modify the estimator process accordingly. Let us review our model assumptions:

1. We assume that the genes directing the secondary pathway are always functioning.

2. We assume that the constants in the gene expression model and secondary control are all constant and remain so.

3. We assume that the A/R gene may be on or off. They are controlled via some tertiary process yet to be determined.

Thus, we can consider the example of a three gene system with two A/R gene per expression gene we have 4X4X4 possible states. This means we have 64 possible color states. If we have n A/R genes per expression gene and we have m expression genes we have 2^{nm} possible color states. Now the above algorithm is a least squares estimate algorithm given an A/R gene state. We now propose a model where we first estimate the state of the A/R genes and then given that state we use the least squares approach to estimate the a values which remain. Thus if a specific A/R gene is in a 0 state we then zero out its effecting a value and estimate the remaining a values as we would have done before. The mathematical analysis to justify the algorithm uses the MAP or maximum a posteriori estimate approach (see [50]). Specifically, we maximize the following:

$$\frac{\partial \ln\ p_{a/z}(A/Z)}{\partial A} = 0 \qquad (29)$$

[50] has shown the equivalence to the minimum mean square estimator (MMSE) approach or the Bayesian analysis. Thus we write the MAP estimator as:

$$\frac{\partial \ln\ p_{a/z_C,z_D}(A/Z_{Continuous},Z_{Discrete})}{\partial A} = 0$$

is; (30)

$$\frac{\partial \ln\ p_{a/z_c} p_{z_c/z_D}}{\partial A}$$

clearly from the above we can separate the two optimizations. Namely this tells us that we can, first estimate the binary values of the on/off states of the A or R genes. Then we can use the standard approach to obtain the continuous, C, elements, in our current case the a values. The best estimator for the binary part is a standard MAP (maximum a priori) estimator using a threshold. We can perform this task by examining each microcell entry for it being active or inactive. The algorithm for the calculation is shown in Figure 7. We perform the functions as we stated above (see [51] for a detailed analysis of this approach. In [51] there is developed an integrated outlier and estimation methodology as applied herein).

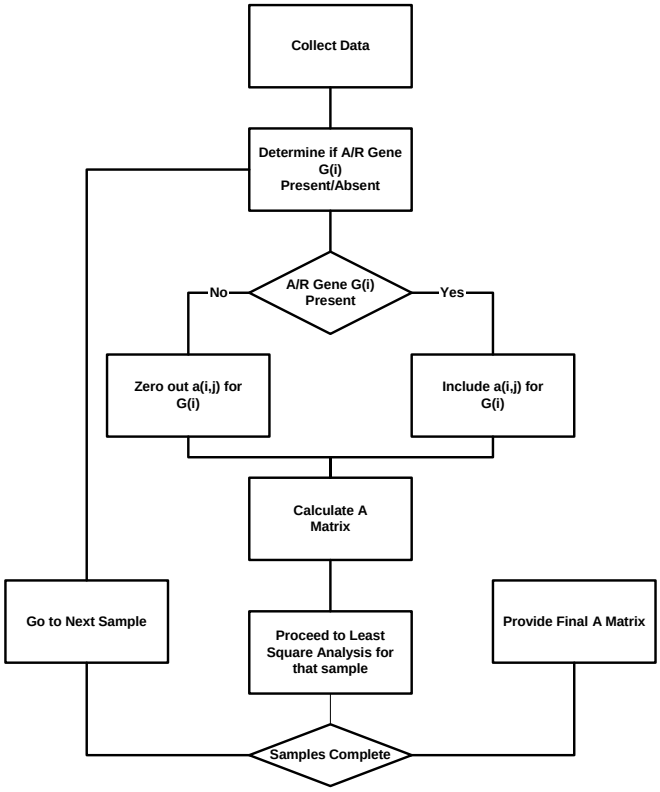

Figure 7: Algorithm for the calculation of the testing and elimination of "zeroed" gene effects.

IV. Estimation versus Identification

The method of estimating the structural elements of the gene expression can be structured using a standard set of methodologies. In particular we use the two approaches used in [43] and [51]. The [43] approach was applied to estimating the constituent chemical concentrations of the upper atmosphere, namely the inversion problem, using transmitted light as the probe mechanism. In this case we seek to estimate the gene expression matrix using the concentrations of secondary chemicals as expressed in color concentrations. This is in many ways a similar problem.

A. The Model

Let us consider a six gene model, two color modifying genes and four control genes, two each. The model is as follows. First is a general linear model for the gene production:

$$\frac{dx(t)}{dt} = Ax(t) + u(t) + n(t) \qquad (31)$$

Then the entries are as follows:

$$A = \begin{bmatrix} a_{11}..a_{12}..a_{13}..0..0..0 \\ 0...a_{22}..0...0...0...0 \\ 0...0...a_{33}..0...0...0 \\ 0...0...0...a_{44}..a_{45}..a_{46} \\ 0...0...0...0...a_{55}..0 \\ 0...0...0...0...0...a_{66} \end{bmatrix}$$

and (32)

$$u(t) = \begin{bmatrix} u_1 \\ ... \\ u_6 \end{bmatrix}$$

and we assume a system noise which is white with the following characteristic:

$$E\left[n(t)\right] = 0$$
$$\textit{and} \quad (33)$$
$$E\left[n(t)n(s)\right] = N_0 I\delta(t-s)$$

Now we can define:

$$A = \begin{bmatrix} A_1...0 \\ 0...A_2 \end{bmatrix} \quad (34)$$

where we have partitioned the matrix into four submatrices. This shows that each gene and its controller are separate. Now we can determine the concentrations of each protein in steady state as follows, neglecting the Gaussian noise element for the time being:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = -A_1^{-1} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}$$
and (35)
$$\begin{bmatrix} x_4 \\ x_5 \\ x_6 \end{bmatrix} = -A_2^{-1} \begin{bmatrix} u_4 \\ u_5 \\ u_6 \end{bmatrix}$$

We will argue that finding either the matrix elements or their inverse relatives is identical. Thus we focus on the inverse elements. Now the concentrations of the anthocyanins are given by the 2 x 2 vector as follows:

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} c_{11}...0...0...0...0...0 \\ 0...0...0...c_{24}...0...0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = Cx \quad (36)$$

The color model remains the same.

B. The Estimator Model

The system model is as follows. Let us begin with a model for the vector a that we seek:

$$\frac{da(t)}{dt} = 0 : where$$
$$a(t) = \begin{bmatrix} a_1 \\ ... \\ a_5 \end{bmatrix} \quad (37)$$

In this case we have assumed a is a 5 x 1 vector but it can be any vector. The measurement system equation is given by:

$$z(t) = g(a,t) + w(t) \quad (38)$$

where z is an m x 1 vector. In this case however we have for the measurement the following:

$$z(t) = \begin{bmatrix} m_1 \\ m_2 \\ m_3 \\ x_1 \\ ... \\ x_6 \end{bmatrix} = g(a,t) + w(t) \quad (39)$$

we now expand in a Taylor series the above g function:

$$g(a,t) = g(a_0,t) + C(a_0,t)\left[a(t) - a_0(t)\right] +$$
$$\frac{1}{2}\sum_{i=1}^{N} \gamma_i \left[a - a_0\right]^T F_i \left[a - a_0\right] + ... \quad (40)$$

where we have:

$$C = \begin{bmatrix} \frac{\partial g_1}{\partial a_1} & ... & \frac{\partial g_1}{\partial a_n} \\ .. & .. & .. \\ \frac{\partial g_m}{\partial a_1} & ... & \frac{\partial g_m}{\partial a_n} \end{bmatrix} \quad (41)$$

Thus we have for the measurement:

$$z(t) = C(t)a(t) + \left[g(a_0) - C(a_0)a_0(t) \right] \quad (42)$$

We now use standard Kalman theory to determine the mean square estimate;

$$\begin{aligned} &\frac{d\hat{a}(t)}{dt} = P(t)C^T(t)K^{-1}(z - C(t)\hat{a}(t)) \\ &where \\ &\frac{dP(t)}{dt} = -P(t)C^T(t)K^{-1}C(t)P(t) + \\ &\sum_{i=1}^{N} PF_i P\gamma_i{}^T K^{-1}(z - g(a_0) \\ &where \\ &K\gamma(t-s) = E\left[w(t)w^T(s) \right] \end{aligned} \quad (43)$$

In discrete time we have the equation:

$$\hat{a}(k+1) = \hat{a}(k) + PCK^{-1}\left[z(k) - \hat{z}(k) \right] \quad (44)$$

Which is identical to the equation we derived from the Newton method.

C. Model Variants

We now want to consider several variants on how color can be generated. In our model we assumed that Expression constants and drivers remained the same throughout but that they were turned on and off thus generating differing colors. However there are many possibilities,

1. Expression Constants Vary: In this case the a values vary from color to color. From a gene expression perspective we cannot find this an attractive alternative. However it may be conceivable that there are secondary controller A/R genes which may be playing a role which may be unidentified at the time of the experimental analysis.

2. Expression Drivers Vary: The u values we have used to be the steady state drivers may be affected by various factors, including local factors such as plant acidity and location. We have assumed we know these variables. This assumption is based upon some past experimental analyses. However these may vary and must be taken into account. To do so we can expand the model to estimate them as well.

3. Expression Constants are on/off: This is assumed in the model we have developed herein. This model assumes that all variables are constant and that we have just an on and off process of A/R genes.

4. Expression Drivers are on/off: This is an intriguing alternative with no known physical embodiment but it may be the case from time to time. However the effect is the same for A/R expression as if we assumed the constants went on or off.

Any combination of the above may also occur.

V. Conclusion

In this paper we have presented an interesting genus to study with respect to gene expression and ultimately the control of gene expression. The phenotypes are quite obvious in flower colors and in addition the hybridizing which has led to a wealth of examples has been done just in the past one hundred years. Also we have a reasonably clear understanding of the underlying species and we can readily assess the complexity of the species DNA structure.

There are three problems for which this approach applies:

1. **Analysis:** In the analysis problem we assume we know the expression dynamics and the secondary production model. Given those two we determine what color we get. We develop this in detail.

2. **Identification:** The identification problem is one in which we know the secondary processes, we have many color samples and we know the protein concentrations which yield each color. Then we ask how do we determine the A matrix for the gene expression?.

3. **Design:** This problem is of significant interest. We seek a desired output state, color in our example. We know the gene expression dynamics and the secondary model. We then ask how do we modify the gene expression model to obtain the desired output?

We also have a well defined and understood set of pathways that give rise to the phenotype. We further know the effecting proteins and enzymes. We also know the gene which affects the proteins in question. Finally we have well accepted models for the expression of the genes and we can use generally accepted models for the dynamics of gene expression.

This has led us in our final section to a modeling of gene expression as a set of definable dynamic systems. We have used a certain set of those systems to discuss examples. However certain key questions remain:

First, what are the dynamic models which can adequately and correctly describe the abrupt coloration of the flowers? We have a good understanding of many of the unstable dynamic systems models which can describe such phenomenon but what is the relationship between what occurs in the cell and what the models describe?

Second, we have used an ensemble approach versus the microbiologists' time approach to modeling the system. We have posited an equivalency based upon the Ergodic Theorem, which states the time average and ensemble average are equal. However there is no experimental proof of this fact.

Third, in any systems approach, we always look at issues as observability and controllability. Observability concerns whether we can see the outputs knowing the system model and can we predict the initial condition. This must be validated experimentally. Controllability is simply can we drive the system to a desired state with a control function. The controllability question goes to the heart of flower color design. If we accept the validity of our models the answer appears to be determinable for any set of defined pathways.

Fourth, we have suggested a microarray approach to estimating the coefficients of the dynamic system. This is one of many possible techniques. The first part we should do is address this from an experimental perspective. Namely perform the microarray analysis. The second part is to investigate alternative methods of solving the system identification problem via alternative bench based validation tools.

Fifth, specific phenotypic design must be considered in more detail and experimentally validated.

Sixth, we use a stochastic model for the expression and pathway analysis. We used this as a way to account for dimensions we could not include because they were expressed at too low a level or because we had no knowledge of their existence. Thus we argued that noise may be true random processes or the aggregation of currently unknown tertiary processes. Experimental validation of this modeling element must be performed.

Seventh, can this approach be carried over to any other cell line? The answer we believe is yes it can and readily. What we have done herein is to focus on phenotypic characteristics and ones which are readily characterizable by well understood pathways. Such systems exist in many other biological systems including the human.

VI. References

[01] Dahlgren, R.M.T., T**he Families of Monocotyledons**, Springer (New York) 1985.

[02] Erhardt, W., **Hemerocalis**, Timber Press (Portland, OR) 1992.

[03] Taiz, L., E. Zeiger, **Plant Physiology**, Benjamin Cummings (Redwood City, CA) 1991.

[04] Munson, R., **Hemerocalis, The Daylily**, Timber Press (Portland, OR) 1989.

[05] Stout, A.B., **Daylilies**, Saga Press (Millwood, NY) 1986.

[06] Dey, P. M., J. B. Harborne, **Plant Biochemistry**, Academic Press (New York) 1997.

[07] Griffiths, A., et al, **Genetic Analysis 5th Ed**, Freeman (New York) 1993.

[08] Mohr, H., P. Schopfer, **Plant Physiology**, Springer (New York) 1995.

[09] Watson, J., et al, **Molecular Biology of the Gene**, Benjamin Cummings (San Francisco) 2004.

[10] Kohane, I., et al, **Microarrays for an Integrative Genomics**, MIT Press (Cambridge) 2003.

[11] Chung, M., J. Noguchi, *Geographic spatial correlation of morphological characters of Hemerocalis middendorfii complex*, Ann Bot Fennici Vol 35, 1998, pp. 183-189.

[12] Chung, M., *Spatial Structure of three Populations of Hemerocalis hakuuensis*, Bot. Bull. Acad. Sci., 2000, Vol 41, pp. 231-236.

[13] Noguchi, J., H. De-yuan, *Multiple origins of the Japanese nocturnal Hemerocalis citrina*, Int Jrl Plant Science, 2004, Vol 16, pp. 219-230.

[14] Tomkins, J. R., *DNA Fingerprinting in Daylilies*, Parts I and II, Daylily Journal, Vol 56 No 2 and 3 2001, pp. 195-200 and pp. 343-347.

[15] Tomkins, J. R., *How much DNA is in a Daylily*, Daylily Journal, Vol 58 No 2 2003, pp. 205-209.

[16] Tomkins, J., et al, *Evaluation of genetic variation in the daylily (Hemerocalis) using AFLP markers*, Theor Appl Genet Vol 102, 2001, pp. 489-496.

[17] Winkel-Shirley, B., *Flavonoid Biosynthesis*, Plant Physiology, Vol 126 June 2001 pp 485-493.

[18] Mol, J, et al, *How Genes Paint Flowers and Seeds, Trends in Plant Science*, Vol 3 June 1998, pp 212-217.

[19] Mol, J., et al, *Novel Colored Plants*, Current Opinion in Biotechnology, Vol 10, 1999, pp 198-201.

[20] Holton, T., E. Cornish, *Genetics and Biochemistry of Anthocyanin Biosynthesis,* The Plant Cell, Vol 7, 1995, pp 1071-1083.

[21] Naik, P. S., et al, *Genetic manipulation of carotenoid pathway in higher plants*, Current Science, Vol 85, No 10, Nov 2003, pp 1423-1430.

[22] Bartel, B., S. Matsuda, *Seeing Red*, Science, Vol 299, 17 Jan 2003, pp 352-353.

[23] Szallasi, Z. **System Modeling in Cellular Biology: From Concepts to Nuts and Bolts**. MIT Press (Cambridge) 2006.

[24] Hatzimanikatis, V., *Dynamical Analysis of Gene Networks Requires Both mRNA and Protein Expression Information*, Metabolic Engr, Vol 1, 1999, pp. 275-281.

[25] Vohradsky, J., *Neural Network Model of Gene Expression*, FASEB Journal, Vol 15, March 2001, pp. 846-854.

[26] Perkins, T., et al, *Inferring Models of Gene Expression Dynamics*, Journal of Theoretical Biology, Vol 230, 2004, pp. 289-299.

[27] Chen, T., et al, *Modeling Gene Expression with Differential Equations*, Pacific Symposium on Biocomputing, 1999 pp. 29-40.

[28] McGarty, T., **Stochastic Systems and State Estimation**, Wiley (New York) 1974.

[29] Turing, A., *The Chemical Basis of Morphogenesis*, Phil Trans Royal Soc London B337 pp 37-72, 19459.

[30] Tinoco, I. et al, **Physical Chemistry**, Prentice Hall (Englewood Cliffs, NJ) 1995.

[31] Goodwin, T.W., **Chemistry and Biochemistry of Plant Pigments**, Vols 1 and 2, Academic Press (New York) 1976.

[32] Atkins, P. **Physical Chemistry**, Freeman (New York) 1990.

[33] Murrell, J., *Understanding Rate of Chemical Reactions*, University of Sussex.

[34] Murray, J., **Mathematical Biology**, Springer (New York) 1989.

[35] Schnell. S, T. Turner, *Reaction Kinetics in Intracellular Environments with Macromolecular Crowding*, Biophys and Molec Bio vol 85 2004 pp. 235-260.

[36] Jaakola, L. et al, *Expression of Genes Involved in Anthocyanin Biosynthesis*, Plant Physiology, Vol 130 Oct 2002, pp 729-739.

[37] McMurry, J., Begley, T., **The Organic Chemistry of Biological Pathways**, Roberts & Company Publishers, 2005.

[38] Campbell, A., L. Heyer, **Genomics, Proteomics, and Bioinformatics**, Benjamin Cummings (New York) 2003.

[39] Lesk, A., **Bioinformatics**, Oxford (New York) 2002.

[40] Judd, D. B. et al, **Color** 2ND Edition, Wiley (New York) 1963.

[41] Levi, L., **Applied Optics**, Wiley (New York) 1968.

[42] Durrett, H., **Color**, Academic Press (New York) 1987.

[43] Tobias, A., *Directed Evolution of Biosynthetic Pathways to Carotenoids with Unnatural Carbon Bonds*, PhD Thesis, Cal Tech, 2006.

[44] Causton, H. et al, **Microarray Gene Expression and Analysis**, Blackwell (Malden, MA) 2003.

[45] Krane, D., M. Raymer, **Bioinformatics**, Benjamin Cummings (New York) 2003.

[46] Athans, M. et al., **System, Networks, & Computation; Multivariable Methods**, McGraw Hill (New York) 1974

[47] McGarty, T. P, *The Structure of the Upper Atmosphere*, IEEE Automatic Control 1971.

[48] Ljung, L. **System Identification**, Prentice Hall (Englewood Cliffs) 1987.

[49] Sage A. P., J. Melsa, **System Identification**, Academic Press (New York) 1971.

[50] Van Trees, H. L., **Detection Estimation and Modulation Theory**, Wiley (New York) 1968.

[51] McGarty, T. P., *Bayesian Outlier Rejection and State Estimation*, IEEE AC 1975, pp. 682-687.

Terrence P. McGarty (M'63-06) is Managing Partner of The Telmarc Group and is also Research Affiliate at MIT, Cambridge, MA. McGarty holds a PhD from MIT in Electrical Engineering and Computer Science. He has been on the faculties of MIT, Columbia University, George Washington University, and Polytechnic Institute. He also has extensive business experience in telecommunications having managed, founded, and directed over a dozen companies in the past forty five years. McGarty also has studied at The New York Botanical Garden receiving Certificates in Botany and Horticulture.