PROJECT DESCRIPTION

1. Introduction

There are a number of situations where complex genetic control systems may be necessary: production of some high-value molecules using metabolic engineering and probing metabolic and genetic function in organisms. Unlike the situation in electronic circuits, where individual elements of a circuit are well characterized and can be easily assembled from a catalogue of parts to perform a particular control operation, there are few choices in the types of genetic control circuits available to execute a particular operation. Furthermore, many genetic control circuits are incompletely characterized so that they do not give expected results when used under a new set of environmental conditions. In order to reliably make such predictions quantitative analysis and robust design must be made for all levels of the control system from the physical properties of individual molecular parts to the heterogeneity of the cellular population behavior. The basis for the design of a particular genetic control system for any type of control strategy is a set of well-characterized genetic "parts" and a protocol for diagnosing the behavior of networks of these part *in vivo* sufficient for rational and directed evolutionary correction of non-compliant behavior

1.1. Goal and Specific Aims

The goal of this project is to develop a theoretical and experimental framework to characterize naturally occurring genetic control circuits and to assemble novel genetic control circuits from the characterized parts to meet a particular control strategy. To this end, the specific aims are:

1. to create a streamlined protocol for rapidly building experimentally validated, dynamical models of biological parts (e.g., protein kinetics, translational control, mRNA degradation, elongation, transcription initiation) of sufficient detail that the behavior of networks composed of interconnected sets of these parts may be robustly predicted and engineered;
2. to create very detailed, experimentally validated models of cellular environmental sensing networks in which there are varying degrees of previous knowledge;
3. to develop a streamlined protocol for computer-aided design of gene expression networks with specified input/output behaviors from the previously measured biological components using these validated network models and the mathematical description of their components.

As model systems, we will study mathematically, computationally and experimentally, three "orthogonal" examples of genetic expression switches in *E. coli*: the chemosensing arabinose promoter system, the type-1C pili phase variation control network, and the OmpR mediated osmoregulatory system.

Achieving these goals requires the use of techniques from molecular biology, analytical biochemistry, statistical data analysis, database design and mathematical modeling and the project brings together a team of biologists, chemists and engineers to achieve these goals. Graduate, undergraduate, and post-doctoral students trained in this effort will be a new generation of 'biological control' engineers to meet the needs of the 21st century biotechnology industry.

1.2. Significance

A genetic "parts catalogue" will allow the scientist or engineer to design genetic control circuits to meet a particular need. In an industrial setting, these new genetic circuits would allow one to induce gene expression for a particular product using a complex control strategy to maximize product formation. For the scientist, the elements of the "parts catalogue" would allow one to develop novel control strategies to probe a complex genetic process.

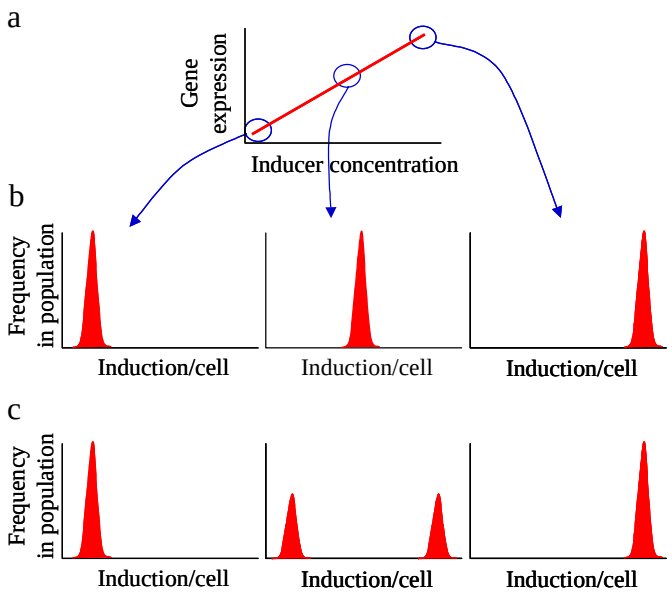

Figure 1. Expression from an inducible promoter as a function of inducer concentration. In general, there is a range of inducer concentrations that results in a linear, population-averaged response from the promoter (a). For most situations, it is desirable that all cells in the culture be induced to a consistent level (b). However, it has been found that at intermediate inducer concentrations a fraction of the population is fully induced and a fraction of the population is uninduced (c). This all-or-none response of the promoter can lead to a heterogeneous population.

One of the primary goals of this project is the design and construction of novel genetic control systems that will allow the use of multiple stimuli to induce and/or repress gene expression along a specified time course. One would like to have a number of different "orthogonal", i.e., non-interfering environmental inputs that allow fine control of a number of different gene products, the expression of which would be contingent on the right constellation of environmental inputs.

2. Background

Gene expression engineering work to date has largely focused on the design of promoters that can be turned ON/OFF by a change in an environmental condition, or inducer. One of the best known genetic control systems is the lactose-inducible *lacZYA* promoter. This genetic control system has been remodeled in several ways for the particular application of interest. However, in general, the overall induction behavior of the promoter does not change. The *lac* operon of *E. coli* exhibits an all-or-none phenomenon, in which the population-averaged expression of a gene under control of the lactose-inducible promoter (P_{lac}) varied roughly linearly with the amount of lactose (30). However, at intermediate concentrations of inducer, P_{lac} is fully induced in a fraction of the cells and uninduced in the remainder of the cells; the fraction of cells fully induced varies with the amount of inducer added to the culture (Figure 1). It is thought that all-or-none gene expression is due to the coupling of the expression of the gene encoding the transporter to the inducer. This all-or-none phenomenon could have deleterious effects on the host, as those cells expressing the genes would be subject to more metabolic burden (and grow more slowly) than those cells not expressing the genes. If the cells are producing a product whose composition depended on the level of inducible gene expression, this all-or-none phenomenon could lead to subpopulations of cells with very different product compositions.

2.1.1. Chemically inducible promoters

Several chemically inducible expression systems are now available for use in bacteria. Besides the versatile and extensively used lactose-inducible promoters and the arabinose-inducible expression system described below, there are expression systems based on the regulatory genes for the upper (3, 4, 11, 13, 31) and *meta*-cleavage (3, 4, 11-13) pathways for toluene degradation from the TOL plasmid and for naphthalene degradation from the NAH plasmid (11, 42) are now available.

2.1.2. Starvation promoters

For many metabolic engineering applications, particularly those related to bioremediation of environmental contaminants, it is desirable to use promoters induced by starvation. The phosphate-starvation promoter of *E. coli* has been widely used for the expression of heterologous

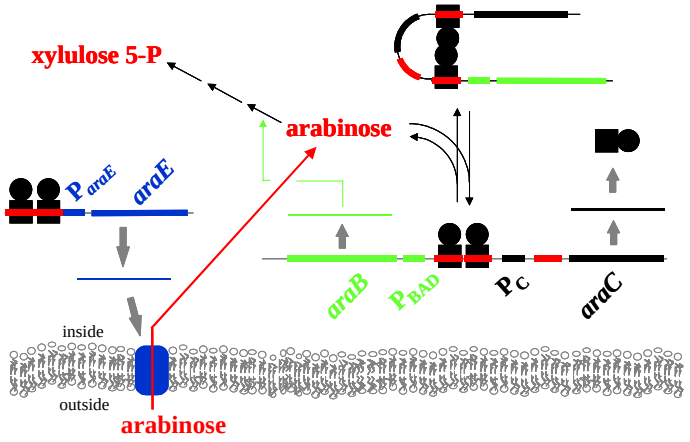

Figure 2. Control of genes responsible for arabinose transport and catabolism.

genes (36). By placing a gene of interest under control of the alkaline phosphatase promoter -- which is induced in the absence of external phosphate -- one can use phosphate starvation to control the expression of heterologous genes during the non-growth phase.

2.1.3. Constitutive promoters

Recently, Jensen and coworkers showed that one could vary the nucleotide sequence in the –10 or –35 region of a promoter, in the 16-bp spacer between these regions, or both and change the strength of that promoter over several orders of magnitude (17, 18). Using this technology, they were able to generate a family of constitutive promoters for use in *E. coli* and *Lactococcus lactis*. These promoters should become extremely important for engineering metabolism when an inducible promoter is not needed.

2.2. Environmentally-switched model systems: design inspiration

As described above, current gene expression system designs are limited in that they are generally of the all-or-none variety, and are one-dimensional. If we are to build more complex expression/repression systems capable of producing a variety of expression patterns (see Figure 7 in Section 4) of a number of genes, simultaneously and at low cost, we must broaden our design repertoire. Since nowhere has this gene circuit engineering been done more elegantly than in cells, we look to a number of naturally occurring gene circuits in E. coli for inspiration. In particular, we study the three following environmentally switched circuits: the chemosensing arabinose promoter system (system 1), the type-1C pili phase variation control network (system 2), and the OmpR mediated osmoregulatory system (system 3). These genetic circuits were chosen because they span a range of control types from those that exhibit an irreversible threshold response to those that have a reversible, fine graded response.

2.2.4. Chemosensing arabinose promoter (System 1)

One of the best-studied chemical induction systems is the arabinose (or ara) operon (reviewed in (34, 35)). Arabinose is transported into the cell via the high-affinity low-capacity transporter AraFGH or via the low-affinity high-capacity transporter AraE (Figure 2). Once inside the cell, the arabinose is transformed into D-xylulose 5-phosphate by AraBAD. All of the genes involved in arabinose transport and catabolism are under control of arabinose-responsive promoters, the activity of which is controlled by the AraC protein. In the absence of arabinose, the AraC dimer binds two different DNA sites some 210-bp apart (facilitated by a DNA loop) and represses transcription from the divergent promoters for *araBAD* and *araC* as well as from the *araE* and *araFGH* promoters. In the presence of arabinose, AraC changes AraC conformation and stimulates transcription from *araBAD* and *araE*. During the transition in conformation, it appears that expression of *araC* is transiently derepressed. AraC also binds D-fucose, a competitive inhibitor of arabinose binding, but D-fucose does not stimulate the change in AraC structure upon binding. A recent study showed that titration of D-fucose could be used to repress expression from P_{BAD}.

Engineered plasmid vectors carrying the *araC-*P_{BAD} fragment from the *ara* operon have been successfully used in *Escherichia coli* and *Salmonella typhimurium* as recombinant gene expression systems (15). This self-regulating system provides fine control of expression, tight repression in the absence of inducer, and induction over a 1000-fold range in the presence of inducer (15). With the development of broad-host-range plasmids containing the *araC-*P_{BAD}

repressor-promoter assemblage (29), this system is now also available for use in non-enteric, Gram-negative bacteria.

Unfortunately, the response of the *araC-*P_{BAD} system in individual cells to arabinose concentration is not linear (Figure 1). In a recent study, Siegele and Hu (37) demonstrated all-or-none behavior of *E. coli* cells containing P_{BAD}::*gfp* constructs exposed to intermediate arabinose concentrations similar to that observed for the *lac* operon by Novick and Weiner (30). This phenomenon was attributed to the inducible L-arabinose transport systems (*araE* and *araFGH*).

2.2.5. Type 1-C pili phase variation control network (System 2)

We propose to study the type-1C piliation switch in *Escherichia coli*, an ideal candidate for control motif identification and re-engineering because of its ability to sense the environment and actuate real-time survival-enhancing responses.

The *fim* switch controls expression of type-1C fimbriae (pili), adhesin-tipped molecules mediating attachment to the mannose-containing receptors found in many host tissue surfaces, and thought to be a virulence factor in urinary tract infections [Schaeffer, 1987 #1; Connell, 1996 #9; Langermann, 2000 #28]. Expression of type-1 pili in E. coli is phase variable, meaning that individual cells in a population randomly alternate between a piliated state, in which pili are built densely on the outside of the cell, and a non-piliated state.

Figure 3 shows the molecular level schematic of the *fim* switch network as currently understood. The phase variation of type 1 pili is controlled by the orientation of a 314 bp length of DNA, an invertible element containing the promoter for *fim* structural subunit genes (*fimA-fimH*) [Abraham, 1985 #6]. In one orientation (ON), this promoter directs transcription of the *fim* structural subunit genes, while in the other (OFF) orientation these genes are not transcribed. Recombinases FimE and FimB can act independently to invert this switch [Klemm, 1986 #14; McClain, 1991 #15; Gally, 1993 #12], and have been shown to bind to four half-sites flanking the switch boundaries (covering intragenic repeats IRL and IRR) [Gally, 1996 #10]. In addition to FimB and FimE, global regulating proteins IHF, Lrp, and H-NS play a supporting role in *fim* switch control. IHF, necessary for any observable switching to occur, binds to two sites along the DNA, one inside the invertible element (IHFI) and one outside (IHFII) [Blomfield, 1997 #8]. Lrp, which serves to increase switching rates in both directions, an effect amplified by the presence of leucine, binds in dimeric form to three sites within the invertible element, while the Lrp-leucine leucine complex binds to only two of the three sites [Roesch, 1998 #5]. Protein H-NS, a temperature-sensitive repressor, represses expression of *fimE*, *fimB*, and *lrp* by binding at or near their promoters and blocking the formation of an RNAP transcription complex [Olsen, 1998 #29; Olsen, 1994 #16; Oshima, 1995 #17]. Genes *lrp*, *ihfAB* and *hNS* are all known to be self-regulating via repressive binding to their own promoters [Ueguchi, 1993 #30]. The expression of *fimE* is orientationally controlled, meaning that it is expressed when the *fim* switch is in the ON position, but is not detectably expressed when *fim* is OFF [Kulasekara, 1999 #7].

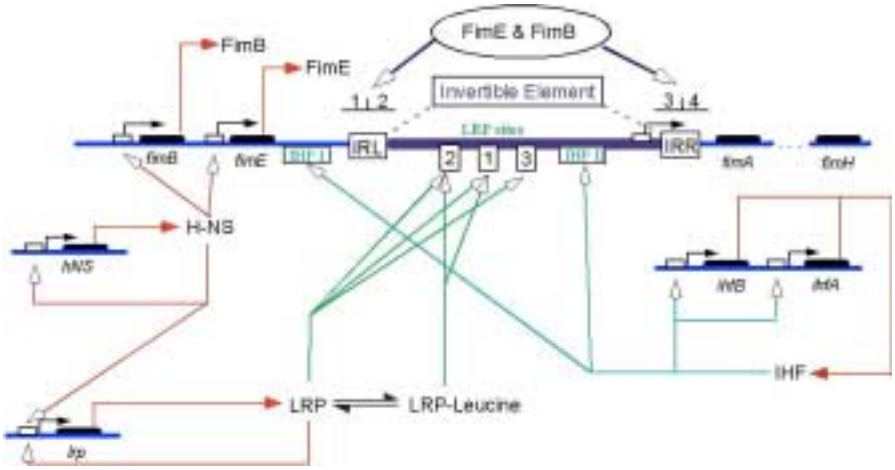

Figure 3. Elements of the Fim system

The *fim* switch control circuitry is responsive to environmental variables such as temperature and the nutritional composition of the medium, and appears to be optimized for

survival either within or outside of a host organism. The ability of this circuit to sense the environment and actuate survival-enhancing changes in switching rates and population heterogeneity (percentage of piliated bacteria in a population) make it an ideal candidate for control motif identification and subsequent re-engineering into novel gene expression/repression systems suitable for laboratory and industrial applications. See Section 3.5 of the proposal for a summary of Wolf and Arkin's preliminary results identifying control mechanisms at work in *fim* circuitry.

2.2.6. OmpR mediated osmoregulation (System 3)

Escherichia coli responds to changes in the osmolarity of its environment by differentially transcribing porin genes *ompF* and *ompC* [Egger, 1997 #25; Pratt, 1995 #23]. At low osmolarity, OmpF is the major porin in the outer membrane. OmpF has a large pore diameter and a fast flow-rate. At high osmolarity, *ompF* expression is repressed and *ompC* is activated. OmpC has a smaller pore diameter and a slower flow-rate [Nikaido, 1993 #24].

The expression of *ompF* and *ompC* is controlled by a two-component regulatory system consisting of the two proteins EnvZ and OmpR. To accomplish this regulation, the sensor EnvZ controls the activity of the transcriptional factor OmpR by (1) responding to environmental stress by converting OmpR from an inactive form to an active protein via phosphorylation [Forst, 1989 #26]and (2) dephosphorylating OmpR-P in the absense of environmental stress [Igo, 1989 #22]. The balance between the kinase and phosphatase activities of EnvZ therefore controls the level of OmpR-P in the cell. Based on the current model, the cellular level of OmpR-P is primarily responsible for the differential expression of *ompF* and *ompC*. A diagram of this relation (from [Lan, 1998 #21; Russo, 1991 #20]) is shown to the left.

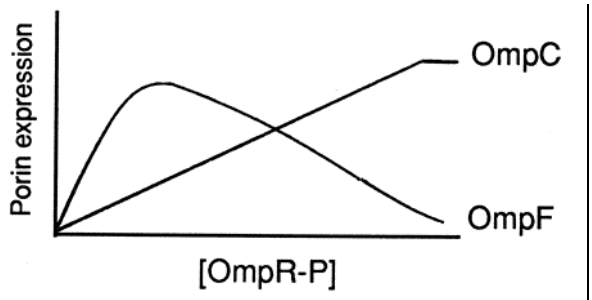

Figure 4. OmpC,OmpF gene expression as a function of phosphorylated OmpR. (This is Figure _1_ from [Lan, 1998 #21])

Notice that OmpC expression first increases and then decreases with increasing OmpR-P concentration (and presumably osmolarity), while OmpF expression increases with OmpR-P concentration to the point of saturation. *As our canonical response curve library (*Figure 7*) includes both regional and linear transfer functions, and the molecular basis for OmpR-mediated osmoregulatory control has been sufficiently well established to construct a detailed network model, this system appears promising for extracting control motifs for industrial re-engineering applications.* An additional advantage to this system is that the controlling environmental variable, osmolarity, is easy and inexpensive to regulate.

2.2.7. Need for modeling, analysis

Environmentally switched Systems 1-3 are marked by complex, nonlinear interactions and the presence of feedback loops. Because of these complexities, mathematical modeling and systems analysis tools are needed to elucidate the 'control motifs' in operation in each system, and their mode of implementation in the network architectures. The isolation of such implementable control motifs is an important step in the design of complex, industrial-grade, orthogonal gene expression/repression systems.

3. Preliminary Work

The Arkin and Keasling laboratories have extensive experience modeling, analyzing, and re-engineering genetic circuitry for industrial purposes. Presented below are a few salient examples involving modeling and analysis of the lac operon, an investigation of the arabinose operator as a model autocatalytic system, design of an arabinose-inducible promoter, modeling

and large-scale system analysis of the λ lysis/lysogeny decision circuit, and modeling and analysis of type 1-C pili phase variation in *E. coli*.

3.1. A mathematical model of the lac operon: inducer exclusion, catabolite repression, and diauxic growth on glucose and lactose.

A mathematical model of the *lac* operon was developed to study diauxic growth on glucose and lactose (41). The model includes catabolite repression, inducer exclusion, lactose hydrolysis to glucose and galactose, and synthesis and degradation of allolactose. In addition to synthesis, degradation and secretion of cAMP were also included in the model. Two models for the phosphorylation of the glucose produced from lactose hydrolysis were also tested: (*i*) phosphorylation by intracellular hexokinase and (*ii*) secretion of glucose and subsequent phosphorylation upon transport back into the cell. The latter model resulted in weak catabolite repression when the glucose produced from lactose was transported out of the cell, whereas the former model showed no catabolite repression during growth on lactose. Parameter sensitivity analysis indicates the importance of key parameters to *lac* operon expression and cell growth: the lactose and allolactose transformation rates by β-galactosidase and the glucose concentrations that affect catabolite repression and inducer exclusion. Large values of the allolactose hydrolysis rate resulted in low concentrations of allolactose, low-level expression of the *lac* operon, and slow growth due to limited import and metabolism of lactose; small values resulted in a high concentration of allolactose, high-level expression of the *lac* operon, and slow growth due to a limiting concentration of Glu6P formed from allolactose. Changes in the rates of all β-galactosidase-catalyzed reactions showed similar behavior, but had more drastic effects on the growth rate. Changes in the glucose concentration that inhibited lactose transport could extend or contract the diauxic growth period during growth in the presence of glucose and lactose. Moreover, changes in the glucose concentration that affected catabolite repression affected the cAMP levels and *lac* operon expression, but had a lesser effect on the growth rate.

3.2. Investigating Autocatalytic Gene Expression Systems Through Mechanistic Modeling.

The model described above was deterministic and did not describe the stochastic nature of gene expression. To more realistically capture the dynamics of gene expression in a population, Carrier & Keasling developed a structured model of gene expression that incorporates the stochastic behavior of cellular processes. This model was used to examine the "all-or-none" phenomenon observed in autocatalytic systems (e.g., the lactose and arabinose operons) (8). Autocatalytic expression systems typically have the genes encoding the inducer transport proteins controlled by internal inducer levels, so that transport of the inducer increases production of the transport protein. The model was able to predict the unique behaviors of autocatalytic expression systems that have been experimentally observed and provided valuable insight into the role of population heterogeneity in these systems. The simulations substantiate the importance of stochastic processes on induction of gene expression in autocatalytic systems. The simulation results show that the all-or-none phenomenon is governed largely by random cellular events, and that population-averaged variations in gene expression are due to changes in the frequency of full gene induction in individual cells rather than to uniform variations in gene expression across the entire population. In addition, the model shows how concentrations of inducer too low to induce expression in uninduced cells can maintain induction in preinduced cultures. A comparison of induction behaviors from an autocatalytic system and a system having constitutive synthesis of the transport protein showed that transport protein levels must be decoupled from inducer control to achieve homogeneous expression of a gene of interest in all cells of a culture.

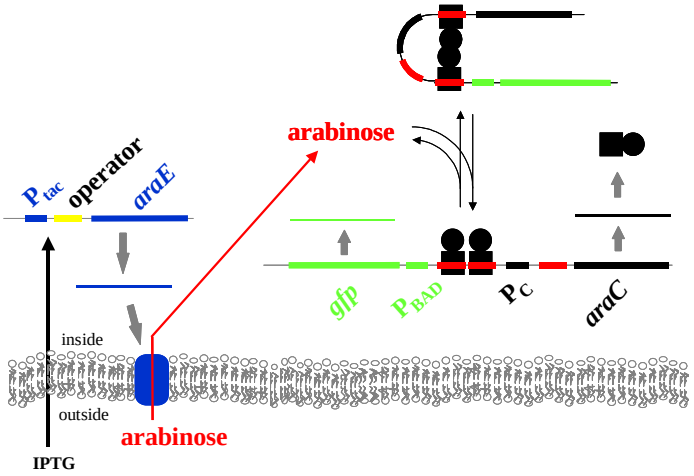

Figure 5. Decoupled transporter/reporter system. The *araE* gene, encoding the arabinose transporter, was placed under control of P_{tac}. The gene encoding green fluorescent protein (*gfp*) was placed under control of the arabinose-inducible promoter (P_{BAD}).

3.3. A regulatable arabinose-inducible gene expression system with consistent control in all cells of a culture

The arabinose-inducible promoter P_{BAD} is subject to all-or-none induction, in which intermediate concentrations of arabinose give rise to subpopulations of cells that are fully induced and uninduced (Figure 1). The modeling work described above indicated that placing the transport gene under inducer-independent control could relieve the all-or-none induction. To construct a host/vector expression system with regulatable control in a homogeneous population of cells, the *araE* gene of *E. coli* was cloned onto an RSF1010-derived plasmid under control of the IPTG inducible P_{tac} promoter (Figure 5) (25). This gene encodes the low-affinity, high-capacity arabinose transport protein and is controlled natively by an arabinose-inducible promoter. To detect the effect of arabinose independent *araE* expression on population homogeneity and cell-specific expression, the *gfp* gene was placed under control of the arabinose-inducible *araBAD* promoter (P_{BAD}) on the pMB1-derived plasmid pBAD24. The transporter and reporter plasmids were transformed into arabinose transport-deficient (*araE* and/or *araFGH*) *E. coli* strains and a commercial, catabolism-deficient (*araBAD*) *E. coli* strain commonly used for heterologous gene expression. The effect of the arabinose concentration and arabinose-independent transport control on population homogeneity were investigated in these strains using flow cytometry. The *araE*$^-$ strains harboring the transporter and reporter plasmids were induced uniformly across the population at all inducer concentrations, and the level of gene expression in individual cells varied with arabinose concentration. In contrast, the strain mutant in only *araFGH*, its corresponding parent, and the catabolism-deficient strain, all harboring the transporter and reporter plasmids, exhibited all-or-none behavior. This work demonstrates the importance of including a transport gene that is controlled independently of the inducer to achieve regulatable

3.4. Modeling & analysis of the lysis/lysogeny switch in λ-phage

The fundamental mathematical modeling software and libraries for genetic and biochemical reaction networks were developed by Arkin, Ross and McAdams for analysis of the λ-phage lysis lysogeny circuits. Though the software is general purpose it was used to model the phage and *Escherichia coli* pathways that govern the two λ-phage decisions (2). They have shown that, in a physically rigorous model of the λ lysis/lysogeny decision circuit, the noise inherent in gene expression (28) can lead to irreducible heterogeneity in this decision; nonetheless, the phage robustly chooses one route or the other. In addition, they tested a number of mechanistic hypotheses about the HflA/HflB mediated degradation of CII and protection of CII by CIII production. The model: 1) Better explained data about percent lysogeny as a function of average phage input than any other model and resolved a conflict with the prevailing wisdom, 2) emphasized the importance of the HflA/HflB proteolytic system as the basis for the lysis/lysogeny switch during the initial infection rather than the OR1-OR3 competitive binding region (this region is certainly more central to the induction process) 3) yielded a prediction for the most likely class of CII/CIII degradative function that was ultimately experimentally proven

correct (in its basic mechanisms). The study again proved that qualitative analysis of even this very simple four-promoter, five-gene system was inadequate for predicting detailed system function and understanding its engineering principles.

3.5. Modeling & analysis of Type 1-C pili phase variation in E. coli

Wolf and Arkin (manuscript in preparation) analyzed the genetic mechanisms responsible for controlling the expression of type 1c fimbriae in *Escherichia coli*, used to attach to mammalian tissues during host colonization and believed to act as virulence factors in urinary tract infections. Because the *fim* gene network (shown in Figure 3) is a stochastic, temperature- and medium- controlled switch designed to maintain survival-enhancing levels of populational heterogeneity both within and outside of its host, it is a good model system for identifying control motifs that allow for adaptation to a rapidly changing environment. By applying tools from mathematical modeling and systems analysis to the now fairly well established molecular basis of the *fim* switch, we determined exactly how the network architecture accomplishes control of the timing, populational heterogeneity and environmental responsiveness of phase variation crucial to the survival of the organism. In particular, we explain (1) the roles of recombinases [FimB], [FimE], their DNA binding affinities and switching rates in observed switching behaviors, (2) why there are two recombinases when one would seem to suffice, (3) the source of the observed ON-to-OFF specificity of FimE, (4) the role of the *fimE* orientational control in switch dynamics, and (5) how temperature tuning of piliation is achieved.

We found that while the percentage of piliated E. coli in a population that has reached steady state depends only on *relative* protein concentrations and parameter values (ratios and differences), the response speed, or amount of time it takes to reach this steady state, depends on *absolute* concentrations and parameter values.

This decoupling means that *response time can be varied without changing the equilibrium %ON*. Any change in environment or growth phase affecting the expression levels of both genes equally will have no effect on %ON but will affect the *amount of time* it takes to reach this equilibrium.

On the question of why the *fim* switch has two recombinases when one would appear to suffice, a single recombinase model does not show the above described degree of decoupling, and, probably more importantly, cannot exhibit the sigmoidal, switch-like control of equilibrium %ON evident in the two-recombinase design.

FimE specificity was found to be primarily a product of differing switching rates of FimB and FimE rather than, as previously hypothesized, differences in DNA binding affinities or the recently discovered orientational control of *fimE*. Orientational control of fimE allows pili turn-ON to go higher and happen faster than it would otherwise.

Our model also gives rise to a plausible hypothesis for an Lrp-mediated mechanism responsible for the observed `temperature tuning' of pili expression to mammalian body temperature. We believe this mechanism to be an example of a larger class of mechanisms responsible for phenotype `tuning' of bacteria to the environment.

Now that many of the engineering design principles of the *fim* circuitry have been elucidated, the next step is to translate these insights into optimal temperature- and medium-controlled expression circuit redesigns to add to our growing gene expression/repression motif library. We expect motifs implementing (1) a temperature-controlled local maximum in expression, (2) a ratio-controlled expression sigmoid with independent control of response strength and response speed, and (3) reversible or irreversibly locked environmentally controlled switches, to be potentially valuable contributions.

4. Research Design

4.1. Overview

The goals of this project are develop a quantitative understanding of naturally occurring gene expression control circuits that may serve as a basis for engineering novel gene expression control systems into cells. The benefits to such research are two-fold: we gain greater insight into the function of complex biological sensing/response systems and we gain the ability to reliably engineer new function into cells. The difficulties involved with these tasks are attributable to three things: (1) the problem of identifying all the direct molecular players and their stoichiometric interactions that provide a given function, (2) understanding the indirect roles of cellular state (i.e., cell growth) on this network, (3) the characterization of the nonlinear, possibly stochastic, *in vivo* kinetics of all these interactions that is the basis of control in these circuits. Advances in genome sequencing and molecular profiling (gene expression, protein separation) have led to great advances in overcoming the first of these problems. These, in turn, have led to a finer ability to create accurate, predictive kinetic models of cellular function. In this work we propose to talk advantage of these advances to create and efficient infrastructure for the analysis and design of gene expression systems. To accomplish this we must develop two protocols, each of which involves iterations of experiment and theory.

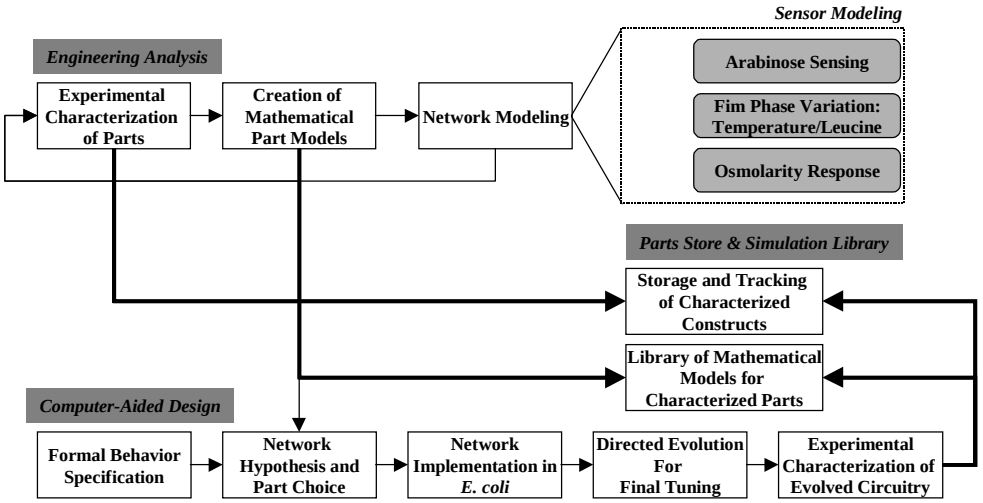

Figure 6. Schematic flow chart of two engineering cycles.

The first protocol is designated the "Engineering Analysis Cycle" (see top half of Figure 6). This cycle is dedicated to created mechanistically detailed kinetic models of gene expression dynamics of a target gene expression circuit. In this work we target three such networks: the arabinose-driven expression system, the control of type-1 pili phase variation and the osmolarity regulatory circuitry. These are three examples of different engineering designs that respond to orthogonal chemical signals. The analysis cycle is designed to: (1) experimentally measure the kinetics and thermodynamics of each of the pieces of these circuits, (2) create detailed models of their individual paet behavior, (3) combine these models into an overall model of circuit function, (4) compare this network model to experimental observations of gene expression, (5) modify the mathematical models of their parts and interactions from this comparison. The part and network models validated during this cycle form the basis for a "parts list" from which new circuits can be built and their behavior predicted.

The second protocol is designated the Computer-Aided Design (CAD) Cycle (see bottom half of Figure 6). The goal of this cycle is to facilitate the design and implementation of novel gene expression control systems meeting arbitrary specifications. Eight such specification types are shown in Figure 7.

The CAD cycle starts with the formal specification of a gene expression profile and its temporal behavior. The CAD program then aids the user in choosing parts and regulatory network structures that fulfill the engineering goals. It does this by drawing on the models described in the Engineering Analysis Cycle that have been stored in the simulation library and on regulatory network theory built into the tool. First, a schematic regulatory network is created that has the appropriate behavior over a wide parameter range and then a particular biological

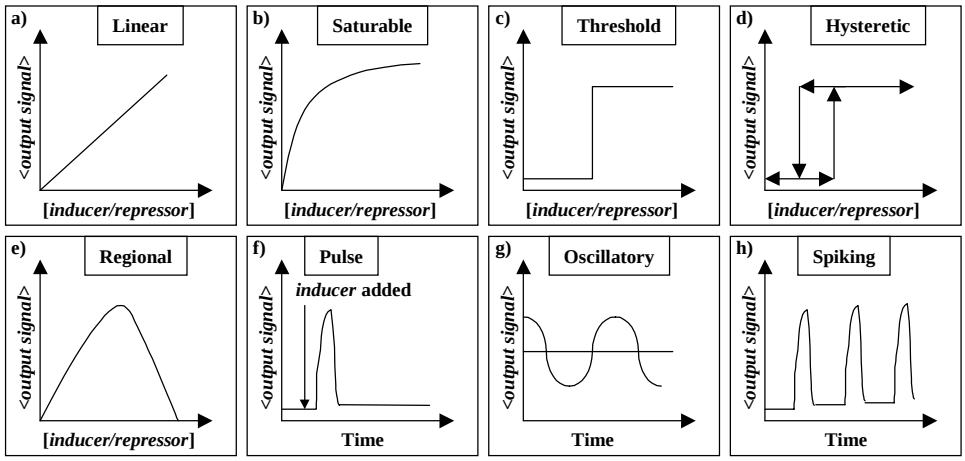

Figure 7. Eight possibilities for desired "induced" gene expression profiles. Note that in a-e the x-axis is the inducer/repressor concentration whereas in f-h, it is time. In fact, both types of gene expression profiles (specifications) have both temporal and signal-level specifications. For example, in case c the time it takes to go from minimum expression to maximum expression might be specified. Other engineering constraints include, for example, specifying the degree of population heterogeneity. These eight types are examples of major classes of control functions used in electrical engineering.

implementation is given and simulated. The CAD tool may or may not be able to meet the exact specifications but will attempt to predict a network that functions in the ball-park of the solution. Once such a circuit is proposed, the appropriate materials are taken from the parts store (or from other sources) and the proposed circuit built in *E. coli* using standard molecular biological techniques. If the circuit function completely fails then a diagnosis cycle is started where in the engineering analysis cycle is used to find the deviations from the prediction. If the circuit function is in the ball-park then rounds of mutagenesis and selection are undertaken to tune the circuit into compliance with the specification where feasible. In this project we will choose three profiles (**a, e,** and **f**) to pursue.

Successful completion of this project will provide not only a detailed understanding of three important gene expression control systems but also create three novel gene expression control circuits for biotechnological use and create a streamlined infrastructure for creation of complex genetic circuitry.

4.2. Model construction, analysis and validation cycle

The goal of the model construction, analysis, and validation cycle of this project is to produce the level of physical description and control system insight necessary to *(1) elucidate the 'control motifs' in operation in each environmentally-switched system under study, (2) uncover their implementation in the network architecture, and (3) design canonical control motifs, to be experimentally validated and placed in a 'motif library' to serve as building blocks in the CAD process (along with an ever-growing set of basic network element/device models).*

4.2.1. Model construction

Our protocol for constructing mathematical models is as follows.

step 1: construct `cartoon' model of genetic network

The first step in developing a mathematical model of a gene network is to construct a graphical representation of the network reflecting the current state of knowledge of the molecular basis of the system in question. An example of such a model is the *fim* switch shown in Figure 3 in Section 2.2.5.

step 2: construct a biophysical model for each class of element and/or device

Because a genetic network is a set of interacting devices, signals, and reactions, construction of a complete system model requires development of a mathematical model of each

network component contributing to gene circuit functioning. Examples of network element/device/reaction classes are transcriptional, translational, and post-translational expression control modules, mRNA degradation, and various enzymatic processes.

step 3: synthesize complete system model from interconnected, instantiated elements/devices

We then instantiate the models according to the circuit diagram and write out the complete set of parameterized equations. Likely forms of the equations describing the complete system model are differential equations, algebraic differential equations, and stochastic differential equations.

step 4: parameterization

We complete the system model by parameterizing it as best we can. Invariably, not all parameters in a model are available in the literature. To determine which of the unknown parameters we need to determine experimentally, and which can be estimated without impacting subsequent analysis and redesign endeavors, we perform a parameter sensitivity analysis, as described below. The 'most vital' parameters are then determined experimentally.

4.2.2. Analysis and simulation of model

The genetic circuit model, once constructed, is then analyzed for its underlying control principles and sensitivity to environmental variables and internal parameters. Tools at our disposal include (1) nonlinear analysis techniques, (2) parameter sensitivity analysis, and (3) numerical simulation**.**

Nonlinear analysis techniques allow one to determine the range of qualitatively different steady state behaviors that can be exhibited by a genetic circuit, and what role each physical system parameter and environmental input plays in the dynamic functioning of a circuit [Guckenheimer, 1997 #19]. AUTO, a bifurcation analysis tool, can aid in this endeavor.

Performance of a ***parameter sensitivity analysis***, aided by prototyping and numerical analysis software like MATLAB, measures how sensitive system behavior is to *small* changes in each system parameter and input. It also produces information on system robustness, and can aid in assigning priorities to missing system parameter values. If the system being studying is large, with many unknown parameters, such prioritization is extremely valuable given limited experimental resources.

The third listed tool, ***numerical simulation***, is a mainstay of genetic circuit analysis. Simulation allows one to explore all aspects of system behavior, and to see how these behaviors change when inputs or parameters are changed or circuit devices are added or removed from the network. Initially, we will use MATLAB and C++ ODE code to simulate the environmentally switched genetic circuits under study. Once Bio/SPICE (Genetic circuit simulation software package under development in Arkin lab) is ready, we will switch to that.

4.2.3. Experimental parts characterization and model verification

The ability to build accurate models for biological control systems depends largely on the quality and quantity of parameter values that are known. While many of the parameter values are known for the systems we have chosen, there are many parameters that have not been measured. Methods to measure some of the most important parameters are described below.

A reference frame. In measuring many of the parameters of gene expression, it is important to consider the frame of reference for the measurements. It is well known that as the environment in which the organism is growing changes, gene expression changes. While the gene expression systems we propose to design may be used under a variety of growth conditions, we will focus on measuring the gene expression parameters under conditions of steady growth (i.e., exponential growth conditions).

RNA synthesis rates. RNA synthesis rates can be measured by inducing expression of the gene of interest. Samples can be taken at various time points after addition of the inducer and instantly added to hot phenol to extract the RNA. The RNA can then be analyzed using Northern blots as described by Carrier and Keasling (5, 7, 9). The increase in the particular RNA

with time gives the RNA synthesis rate. Extrapolating back to the y-axis gives the total time it takes to make the first RNA.

Transcript stability. Stability of the mRNA will be determined using Northern blots as described by Carrier and Keasling (5, 7, 9).

Protein synthesis rates. One of the best methods to measure the synthesis rates of individual proteins is by 2-dimensional gel electrophoresis (1). One must first know the location on the gel of the protein of interest. Once this is known, one can induce the expression of a gene of interest in cells growing exponentially in medium containing a labeled amino acid (i.e., [^{35}S]-methionine). Total proteins are then extracted and electrophoresed, and the gels are exposed to film. The increase in the intensity of the spot corresponding to a particular protein with time is the rate of synthesis of that protein.

Peptide chain elongation rate. The rate of peptide elongation will be determined by measuring the time it takes to make the first protein of interest after induction (1, 27). If the gene of interest were replaced with *lacZ*, then it would be relatively simple to calculate the peptide chain elongation rate. One would expect a parabolic (x^2) increase in β -gal activity with time because mRNA increases with time and β-gal increases with time and mRNA concentration. Thus, plotting the square root of the enzyme activity as a function of time, the x-intercept is the average time necessary to synthesize a protein. Dividing the number of amino acids in the protein by the time necessary to synthesize the protein gives the peptide chain elongation rate.

Protein stability. If the protein is easily assayable, one can induce gene expression. After a significant amount of the enzyme has been synthesized, one can stop transcription and translation using the appropriate antibiotics and measure the decay in enzyme activity. If the protein is not easily assayable, then it will be necessary to have a monoclonal (or possibly a polyclonal) antibody to the protein. By taking samples at specific time points following addition of the antibiotics and then performing Western blots on the samples, one could determine the amount of protein remaining at each time point. Alternatively, one can grow cells in medium containing a labeled amino acid (i.e., [^{35}S]-methionine), stop transcription and translation using the appropriate antibiotics, and determine the rate of protein decay using 2-D gel analysis.

Population homogeneity. An important aspect of many chemically induced systems is population heterogeneity. To determine if a particular promoter is subject to all-or-none expression, one can place the gene for the green fluorescent protein (*gfp*) under the control of the promoter of interest. After inducing the promoter, one can measure the fluorescence in individual cells using flow cytommetry. If a particular promoter is subject to all-or-none expression, the culture should have two subpopulations of cells, one subpopulation will be fully induced and the other will be uninduced.

4.3. Parts store and simulation library

In all other engineering fields design and implementation of a particular construct is aided by having a store of well-characterized standard parts from which the designer may draw. For projects of any complexity the engineer will first use a modeling program (computer aided design tool) to test his or her ideas before implementing them. This is possible because the physical models of the design components are good enough to predict their behavior when connected with other objects and the environment. Often times, the materials are of sufficient complexity that all the engineer can do is set the formal specification of the behavior of their final project, and the program searches through the various part types, sub-element designs, and full implementations until an optimal design in found. The previous section gives a plan for measurement of the physical characteristics of biological components suitable for creation of a "biocomponent" library to form the basis for biological computer-aided design. In addition, the analysis of the sensor networks above will provide an understanding of control motifs. That is, these three circuits are representative of a class of regulatory structures that lead to particular gene expression profiles. What the mathematical analysis yields is a "regulatory template" that identifies the salient features of the circuitry necessary for a particular profile. A different molecular implementation of this template should still yield the same expression profile. These

templates are much like blue prints for a chair. You can implement the chair in oak or mahogany but the wooden pieces must be placed in the correct arrangement for the chair to function.

In order for a computer-aided design tool to be useful, both generic and specific object models must be available and these models must be readily related to material from which the biological circuitry can be built. Thus, the components we propose to measure, chosen initially from the set necessary for creation and validation of the three model systems discussed above, will form the basis set for this mathematical and material library. The specific mathematical models for these parts will be derivative of generic models of transcription initiation, elongation, translation initiation and protein synthesis, mRNA degradation, etc., already defined for other model systems (see *Background Section* above). These generic models will be extended where necessary and parameterized from the measurements above. The resultant specific models will be added to a biocomponent library that will be part of the Bio/Spice simulation and analysis tool described below.

The individual genetic parts described by these models, the promoters, Shine-Delgarno sequences, termination sites, etc, will be placed in plasmid vectors so that they may be easily taken out and inserted in other constructs using standard molecular biological methods. These then serve the role of the material engineering components described above. In order to create behavior (say one of the eight described above) the researcher enters the desired physical characteristics into the CAD tool and, from this specification and the object library, the tool predicts a component list and network structure that produces this behavior. The researcher can then go to the freezer, pull out the appropriate component constructs, and recombine them into the proposed network in order to test the prediction. Though this last experimental part is the most difficult, if the models in the tool are good enough, then the predictions should be robust. The art will be in designing vectors to minimize the work necessary during the network construction phase.

4.4. Computer-aided design cycle

Assuming that the tasks in sections 4.2-4.3 are completed it now becomes possible to implement the computer-aided design cycle described in the Research Design Overview (4.1). This involves formally specifying a target gene expression profile, deriving a feasible regulatory structure for accomplishing this profile, proposing a molecular implementation of this regulatory structure, building the circuit *in vivo* and testing it for compliance with the original specification. If the circuit is compliant with loose constraints, further compliance can be achieved through rounds of directed evolution. The following sections describe each of these tasks.

4.4.1. Formal specification

In the formal specification task we choose a gene expression profile. Examples of such profiles are shown in Figure 7 above. Recently, Gardner et al. [Gardner, 2000 #988], showed the design of a circuit that produced an expression profile like that in Figure 7d and Elowitz, *et al.*[Elowitz, 2000 #983] have produced a system with a profile like that in Figure 7g. These were impressive feats in complex circuit design and showed the power of starting with a regulatory template first in order to get a particular behavior. There has been similar work on creating sharp expression thresholds by Chen, *et al.* [Chen, 1993 #990] and Sektas, *et. al.* [Sektas, 1998 #991]. We, therefore, propose to create circuitry for achieve profiles Figure 7a, e and f. These are particular profiles are of interest from a biotechnological standpoint. The linear response curve (Figure 7a) ensures a stable, predictable response to inducer. The “regional” response curve is useful if the organism to express only under well-defined culture conditions, in a particular place in a chemical gradient or during a particular time in a larger process wherein **a** the inducer is a process product that only transiently passes through the “active” region. The pulse profile is especially useful when the gene expression network is to be used for probing other biological circuitry. Pulse analysis is a standard control analysis in other engineering disciplines (16).

The formal specification of these profiles are fed to the computer as a graph of the major response (much like that shown in Figure 7) as well as a series of constraints on population

heterogeneity, temporal response, maximum and minimum expression levels, etc. These are the data to which models of proposed regulatory networks will be fit.

4.4.2. Network hypothesis and part choice

Once a particular specification has been there begins a round of network hypothesis. In this section, the researcher either specifies a network that he or she believes will accomplish the task and the CAD tool simulates it under different sets of reaction models and parameters sets to fit the specification or network hypotheses are automatically generated from previously stored networks and/or a genetic algorithm approach modified from (14) for building chemical networks to specification. This procedure both tries first to use fully parameterized part models from the validated simulation library and, if these are not sufficient, will use more generic models to accomplish the task. If generic models are necessary for the task then usually, naturally occurring biological parts will need to be found in order to implement the circuit. In some case, it will be possible to re-engineer the circuitry for a particular abstract function. The networks will be chosen not just so that they meet the specification but also for their insensitivity to parameter values.

4.4.3. Network implementation in *E. coli*

The network implementation is the realization of the proposed circuits from section 4.4.2 in *E. coli* using pieces of previously quantitatively characterized circuitry. In addition to the networks we proposed to characterize above, the database current holds data from a number of biosynthetic operons, the lac system, and the λ-phage lysis/lysogeny circuitry material from which is readily available. Implementation is accomplished using standard molecular biological techniques. Once implemented, the behavior of the circuit is characterized using the Engineering Analysis Cycle (Figure 6). Circuits that fail completely will be diagnosed by comparing the experimentally measured behavior to variations in the model behavior.

4.4.4. Directed evolution for fine tuning

It is likely that the implemented network will deviate (sometimes significantly) from the calculated behavior of the network. This is because the artificially constructed interactions and the relationship to the rest of the cellular physiology will not have been fully characterized before. In the cases where there is not complete failure we can begin directed evolution trials in which a mutagen such as EMS is added to the population is grown then screened for the appropriate behavior. For example, for the “regional” control profile a genetically diverse population is exposed to the inducer at low concentration and those members having low expression of a marker are sorted into a pool. This pool is then exposed to the “peak” inducer concentration and those individuals with high expression are sorted into a pool. This last pool is then exposed to even higher concentrations of “inducer” and only those members with low expression are sorted into a pool. The remaining bacteria are then tested over the entire range to make sure the selection worked. Iterated rounds of this procedure may be performed during which the sorting becomes more and more strict. The finally chosen strain is then sequenced in the region of the artificial network and it is determined whether compliance was granted by changes in the circuitry itself or in the broader host physiology.

5. Timeline

The time line for this project is as follows:

Year 1: Construction of initial models for the three natural systems. Quatitative measurement of expression phenotypes. Design and implementation of constructs to measure isolated part properties.

Year 2: Quantitative measurement of transcription initiation, mRNA degradation, protein expression/degradation, etc. for the three systems under study. Comparison of data to models of parts and network. Engineering analysis of network control. Construction of material parts

library and simulation library. Initial network hypothesis generation for the three proposed expression profiles.

Year 3: Implementation of proposed circuitry. Quantitative measurement of circuit function. Diagnosis of failed design and reengineering based on results. Directed evolution of circuits in minimal compliance in order to generate fully compliant circuits. Analysis of evolved circuitry.

6. Work funded by previous NSF grants

6.1. BES-9409603, Design of Auxiliary Chromosomes for *Escherichia coli*

P.I. Jay D. Keasling
Amount $98,039
Period 7/15/94 - 6/30/97

Results: In conjunction with this grant, we developed a low-copy, segregationally stable plasmid for the expression of one or more genes (5, 19). The low-copy expression vector was constructed from a 9-kb region of the *Escherichia coli* F plasmid that contains the *oriV* and *oriS* origins of replication. This plasmid carries the β-lactamase gene to confer resistance to ampicillin, the *araBAD* promoter and the *araC* gene for arabinose-inducible gene expression, and transcription terminators. These low-copy plasmid vectors are extremely stable in continuous culture in the absence of any selection pressure, even during the expression of a heterologous gene. In addition, the low copy number of the heterologous gene allows for very low basal expression in the absence of inducer and low metabolic burden under maximal induction.

Publications (complete citations listed in the References section): Jones & Keasling (19). Keasling, Kuo & Vahanian (23). Kuo & Keasling (26). Wong, Gladney, & Keasling (41).

Human Resources: *Kristala Jones*, PhD, 1999, currently at Merck. *Henry Kuo*, BS, 1996, currently at Chevron. *Robert E. Pape*, MS, 1996, currently at Systemix, Palo Alto, CA. *Gizette Vahanian*, BS, 1993, currently at Boston University. *Patrick Wong*, BS, 1996, currently at U.S.D.A., Albany, CA. *Stephanie Gladney*, BS, 1996, currently at Intel, Santa Clara, CA.

6.2. BES-9502495, CAREER: Strategies for Metabolic Engineering of Bacteria: Novel Synthesis of Biodegradable Polymers

P.I. Jay D. Keasling
Amount $148,273
Period 07/01/95 - 06/30/99

Results: The goal of the work supported by this grant was the development of tools for metabolic engineering of microorganisms. (1) We designed several synthetic mRNA hairpins that allow us to vary the half-life of a transcript over a 10-fold range (5-7, 9, 10). (2) To alleviate the effects of intracellular polyphosphate on the phosphate-starvation promoter, we constructed an *E. coli* strain with the genes encoding polyphosphate kinase (*ppk*) and polyphosphatase (*ppx*) inactivated (38, 39). This new strain allows maximal and predictable control of the phosphate starvation response and, thus, the phosphate starvation promoter. (3) To alleviate the all-or-none phenomenon found for arabinose-inducible promoters, the gene for the inducer transport protein (*araE*) was placed under control of a constitutive promoter (25). In this case, all cells in the culture are induced to approximately the same level, and the level of induction, rather than the fraction of the population that is fully induced, varies with the amount of inducer added. (4) To predict how the heterologous pathways should be balanced with the pathways necessary for growth, we have developed a steady-state mathematical model to predict fluxes through the various metabolic pathways

(32, 33). This model was formulated from the known stoichiometry of the metabolic pathways in bacteria and was solved using linear programming.

Publications (complete citations listed in the References section): Carrier, Jones, & Keasling (5). Carrier & Keasling (6). Carrier & Keasling (10). Carrier & Keasling (7). Carrier & Keasling (8). Carrier & Keasling (9). Carrier & Keasling (8). Keasling, (20). Keasling, Carrier, Jones, Pramanik, & Van Dien (22). Keasling, Carrier, Jones, Pramanik, & Van Dien (21). Keasling, J. D., S. J. Van Dien, & J. Pramanik (24). Khlebnikov, & Keasling (25). Pramanik & Keasling (32). Pramanik & Keasling (33). Van Dien & Keasling (39). Van Dien, Keyhani, Yang, & Keasling (40).

Human Resources: *Trent Carrier,* PhD, 1998, currently at Merck & Co. *Artem Khlebnikov*, Post-doctoral student, 1998 to present. *Jaya Pramanik,* PhD, 1997, currently at IBM, San Jose, CA. *Christina Smolke,* Chemical Engineering PhD student. *Steve Van Dien,* PhD, 1998, currently a post-doc in Spain.

7. Management Plan

This proposal contains three key personnel, Adam Arkin, Jay Keasling, and Denise Wolf, two post-doctoral researchers and two graduate students. We break the project into three broad overlapping areas: Experimental Engineering Analysis, Computational Analysis of Networks, Computer-Aided Design Tools. For the purposes of management, Jay Keasling will be team leader for the first of these, Denise Wolf, the second, and Adam Arkin the third. However, because of the close proximity of these researchers and the tight collaboration necessary for this project this management team will function more as a steering committee than as discrete entities.

Tasks assigned to the Engineering Analysis head are management of the set up and quality assurance of the experimental equipment, development and performance of experimental protocols, development of statistical error models for the data, comparison of models and experimental data.

Tasks assigned to the Analysis of Networks team is the development of the literature database for the model systems, development of the mathematical tools for simulation and analysis of biological network models and comparison to data, development of experimentally validatable models of the three models systems, oversight of the transfer of laboratory data into the central database, analysis of proposed models from the Computer-Aided Design Team.

The CAD team will focus on the creation of an easy formal specification language and network design and hypothesis generation tool, will aid in the creation of the parts database and in the design of the material parts store, aid in developing the molecular evolution techniques (along with the experimental engineering analysis team) for tuning of the networks and for developing the programs for taking experimental measurement of implemented circuits and diagnosing the differences between predicted and observed behaviors.

In order to facilitate group interactions a project web-site will be set up with appropriate database forms, discussion forums, software postings, and news. This site will be broken into public and private sections. The public form will serve published data and models to the biological community and will be one of the forms of dissemination along with publication and speaking.

There will be a super-group meeting once a month where presentations are made from each of the Area teams and formal plans for the next month's projects are made. Area teams will have staggered formal meetings on a biweekly basis. Day to day interactions among the group will be frequent as well.

8. Educational Activities

In the past, experimentalists and 'modelers' were not trained in the same laboratory or even communicated for that matter. Thus, many 'modelers' do not appreciate the intricacies of experiments, and many experimentalists do not appreciate the power of mathematical models.

To bridge this gap, we propose an interdisciplinary training program among three laboratories. While each student will have his/her own individual project, each student will be exposed to all of the aspects of this project. The students trained in this proposed program will gain a broad experience in mathematical modeling of genetic control systems and of biochemical systems in general, in experimentally characterizing key parameters that affect genetic control, and in designing new genetic control systems for practical application. To foster collaborative and interdisciplinary research, we propose the following aspect to the training program.

8.1. Group Meetings

We will have monthly joint group meetings that will serve both as tutorials for modeling and experimentation and as research progress reports.

8.2. Interdisciplinary projects

Graduate and post-doctoral students will be encouraged to take on projects that will have both experimental and modeling components. While a student's project may be predominantly modeling or experimentation, that person will have significant exposure to the other component. For example, a student who is modeling a particular genetic control system would also be involved in experimentally determining some key parameters for the model that may not be available in the literature.

8.3. Expected accomplishments for individual trainees

Trainees will be expected to learn basic techniques in all areas of the program. Trainees will be expected to publish the results of their work in peer-reviewed journals and to present their work at least once at a national/international meeting.