**Optimization of Synthetic Genetic Circuits Using Bacterial Computers**

**Rebecca Evans**

Using bacteria as a computational computer, while an interesting idea, poses many complications. Cells have the capacity to store huge amounts of information in their DNA (Haynes, 2008), but it can prove complicated to manipulate this information. Sometimes the parts act in unpredictable ways. When using *E. coli* to construct a bacterial hash function, Pearson *et al*. found the LuxR promoter to be bi-directional, initiating transcription in both directions, in certain circumstances. This was an interesting discovery, but the bi-directional nature of the promoter made it unusable for their computer. These unpredictable elements make it difficult to engineer a bacterial computer that rivals a silicon computer’s ability to compute.

Also, we know that it is difficult to control gene networks. Liu *et al.* (2011) found that gene regulatory networks had more “driver nodes” than other studied networks such as social networks and the World Wide Web. Driver nodes are nodes in a network that must be controlled in order to control the entire network. They found that most of the nodes in a gene regulatory network must be controlled in order to control the genes and their outputs. In many ways this makes sense because living organisms must be able to respond to small stimuli in their environment. If they had fewer driver nodes in their gene regulatory networks, they would not be able to make fine-tuned adjustments to multiple stimuli. However, this fact makes it difficult for us to control the output of a gene network.

On the flip side of this problem, it is often difficult to model and engineered a genetic circuit’s behavior because of its unpredictable nature. Though we have information about the individual parts (the rate of transcription for promoters, the binding efficiencies of RBS, etc.), we don’t necessarily know how they work in conjunction with each other and within the host cell. It would be especially difficult to model a circuit for fitness. In other words, it is difficult to determine, using computer modeling, what parts working together will yield the optimal output from a circuit to maximize the fitness of the cell.

Thought it is difficult to model a genetic network to figure out what output will optimize fitness, cells figure this kind of problem out every day; they scan genetic circuits and give an output. The cells (by way of natural selection) then determine which genetic makeup best maximizes their fitness. Because it is difficult to model or control a system, we propose allowing the system to control and model itself. We can shuffle different parts into a gene network and allow the cells to give an output, which will affect their fitness under certain selective pressures.

**Proposed Research**

We propose looking at a four-step gene expression cascade. We will provide one precursor substance, which will go through several intermediate steps mediated by enzymes encoded in the cascade, to produce a product.

We will look at four different dimensions within each gene cassette that affect the final output of the circuit: promoters, ribosomal binding sites, degradation tags, and the order/orientation of each gene expression cassette. We can think of the output from the metabolic pathway as a function of the different dimensions, and fitness as a function of the output. So the resulting outputs can be graphed versus the different dimensions.

If we just use four promoters, four ribosomal binding sites, and four degradation tags (we could use more), and taking all the different orders and orientations of the cassettes into account, we have over 6 billion possible combinations. This vast number of combinations makes it impossible for us to individually design each network. For this reason we want to engineer the cells to assemble the combinations of parts and orders, produce an output, and allow us to select for fitness.

**The Pathway**

We had to find a suitable pathway having certain characteristics. First, the final product of this pathway must allow for selection. We can achieve this by using an initial input that is unusable, which, by the enzymes encoded in the pathway, becomes a useable energy source. By only providing the unusable initial input as a food source, the cells that have the optimal pathway will best be able to use the food and should therefore grow the best. For this to work, the pathway must be isolated. This means that the genes in each cassette must be non-native to the *Escherichia coli K-12 MG1655* strain. Also, the initial input and the intermediate products must only be used in this pathway, or only be used in pathways likewise isolated (surrounded by enzymes not native to our strain of *E. coli*). Also, it would be preferable for the initial input to be readily available and inexpensive.

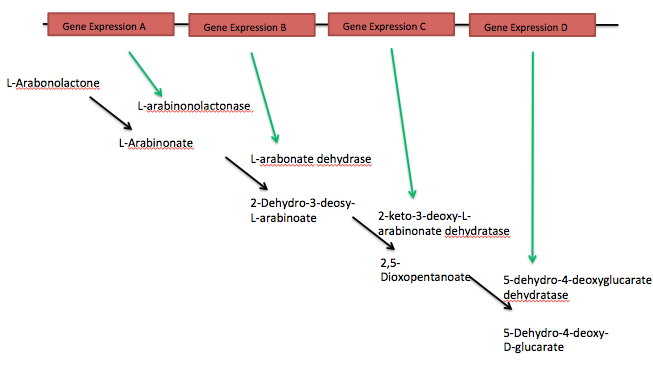
 One possible pathway involves the metabolism of a form of arabinose, an unusable substance for *Escherichia coli K-12 MG1655*, to make a precursor to pyruvate, which would provide energy. The initial input would be L-Arabonolactone, a derivative of the sugar arabinose. This input is ideal because it is only used in this pathway and is surrounded by non-native enzymes.

Figure 1. Information for each enzyme and product found at the [KEGG pathway database](http://www.genome.jp/kegg/pathway/map/map00053.html).

The metabolism of this input is shown in Figure 1. One possible gene sequence for the L-arabinonolactonase enzyme is the [Rpic\_2994](http://www.genome.jp/dbget-bin/www_bget?rpi:Rpic_2994) gene. It has a 915 base pair sequence. One possible gene sequence for the L-arabonate dehydrase enzyme is the [Rpic\_3005](http://www.genome.jp/dbget-bin/www_bget?rpi:Rpic_3005) gene. It has a 1737 base pair sequence. One possible gene sequence for the 2-keto-3-deoxy-L-arabinonate dehydratase enzyme is the [Rpic\_3006](http://www.genome.jp/dbget-bin/www_bget?rpi:Rpic_3006) gene. It has a 942 base pair sequence. One possible gene sequence for the 5-Dehydro-4-deoxy-D-glucarate enzyme is the [Rpic\_4452](http://www.genome.jp/dbget-bin/www_bget?rpi:Rpic_4452) gene. It has a 939 base pair sequence. All of these genes come from the Ralstonia pickettii bacteria.

This acid 5-Dehydro-4-deoxy-D-glucarate would be our final product. Though it, in itself, is not an energy source, the enzyme 2-dehydro-3-deoxyglucarate aldolase, found natively in our strain of *E. coli*, catalyzes the acid to make pyruvate, which is used in glycolysis. If the pathway works as designed, when we only give the cells L-Arabonolactone as an energy source, the cells with the optimal pathway will have the greatest supply of pyruvate and grow the best. Some preliminary experimentation will need to be done to make sure this pathway supplies enough energy to allow the cells to live and grow.

It may be difficult to use L-Arabonolactone as this initial input, because, though it is ideal for our construct, it is not readily available and may be expensive to purchase. It may be possible to use L-arabinose instead as the initial input. This is sugar directly before L-Arabonolactone in the metabolic pathway. L-arabinose is easily accessible and inexpensive. This would involve using a five-step pathway, including the enzyme L-arabinose 1- dehydrogenase as the first step. One possible gene sequence for this enzyme is the [Rpic\_3007](http://www.genome.jp/dbget-bin/www_bget?rpi:Rpic_3007) gene, which also comes form the Ralstonia pickettii bacteria. It has a 939 base pair sequence. The only potential problem with using L-arabinose is that it is also used in the [Pentose and Glucuronate Interconversions](http://www.genome.jp/kegg-bin/show_pathway?org_name=eco&mapno=00040&mapscale=&show_description=hide) pathway and is used by enzymes native to the *Escherichia coli K-12 MG1655* strain. This may however not be a problem because it does not appear to be closely linked to any energy metabolism.

**The first dimension: Promoters**

Our first dimension involves the promoter sequences. We have two different categories of promoters we can use. The first includes inducible promoters such as Plac, Ptet, and PA1lac0. Using inducible promoters would allow us to control when the genes are transcribed. This may be beneficial for conserving the energy necessary for the cells. However, adding chemical inducers could be problematic if they interact with different metabolic pathways or if one was a useable energy source. A second category includes constitutive promoters. Constitutive promoters, because they are always turned on, are always inducing the transcription of the gene. This probably would not be a problem because, until we add our initial input, the enzymes would not have any substance to catalyze, so they would not do much. Constitutive promoters for our set can be picked from a family of constitutive promoters designed by UC Berkley (Parts [J23100 through J23119](http://partsregistry.org/wiki/index.php?title=Part:BBa_J23101)). This family contains 20 different promoters with different strengths. They are each 35 base pairs long.

It may also be beneficial to run the experiment with a set of inducible promoters and a set of constitutive promoters. Since this experiment is a proof of concept, determining if it is possible to optimize a metabolic network for fitness, if it were used in later experiments to determine an optimal pathway, it would be beneficial to be able to run the experiment with both kinds of promoters.

**The second dimension: Ribosomal Binding Sites**

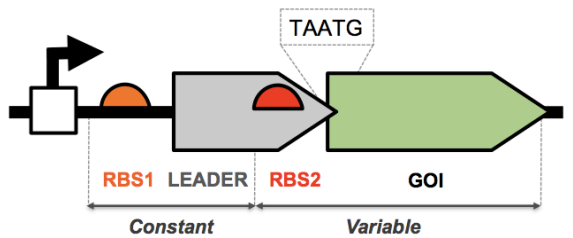
 The second dimension involves using ribosomal binding sites within the C. dog construct. The term “C. dog” refers to the central dogma of the information transfer of DNA to mRNA to protein. This construct is a way to put the ribosomal binding sites together to allow for more “precise and reliable gene expression” (Mutalik, upublished). In this construct (Figure 2), one RBS is kept constant and is placed directly before a leader sequence and a second RBS. Between the second RBS and the sequence for the gene of interest, there is a short TAATG sequence, a merging of a stop codon and another start codon. When this construct is being used, a ribosome lands at the first RBS and encodes the section up to the stop codon (TAA), which is in its open reading frame. The ribosome then disconnects and this short polypeptide is discarded. Then a ribosome can land on the second RBS and begin encoding the gene of interest beginning with the start codon, which is in its open reading frame. The first RBS and the leader sequence always remain the same, but the second RBS can be varied.

Figure 2. Unpublished BIOFAB data of the *E. coli* “C. dog” v1 Bicistronic Translation Junction. Courtesy of Dr. Vivek Mutalik.

When this construct was used, the RBSs gave fairly uniform level of expression across a range of genes of interest, especially when compared with the spread of gene expression resulting when the C. dog construct was not used. By using this construct, we can choose RBSs from those that gave a low level of expression, a medium level of expression, and a high level of expression.

**The third dimension: Degradation tags**

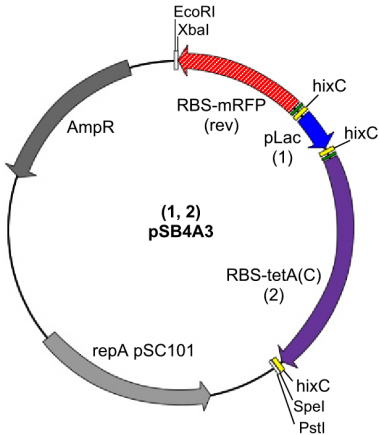
The amino acid residues at the end of a protein often determine the rate at which the protein is degraded. Mogk *et al*. (2007) showed that the amino acid residues at the nitrogen terminal of the polypeptide determine the stability of the entire polypeptide. They called this the N-end rule. From this we know that by altering the end of the protein, we can alter how long it sticks around in the cell. This amount of time can affect how the pathway. Andersen *et al*. (1998) found that the addition of certain sequences to the C-terminal of Gfp gave varying half-lives for the resulting polypeptides, from 40 minutes to several hours. The additional sequences made the proteins susceptible to the proteases present in *E. coli*.

These tags make up our third dimension. Andersen *et al*. studied four different tags and charted the degradation of Gfp in *E. coli*. Each tag is 11 amino acids long (33 base pairs). Two of the four tags led to similar rates of degradation, but the other two led to very different rates. It may also be beneficial to also have, in our degradation tag set, a tag that doesn’t actually lead to degradation (a blank) because we don’t know how degradation will affect the optimization of the network.

**The fourth dimension: Order and Orientation of Genes**

In order to change the order and orientation of the genes, we need a way to flip individual and sections of gene cassettes. There are several protein/base pair sequence systems that can re-orientate segments of DNA. Haynes *et al*. (2008) showed how the Hin/*hix* system could invert a segment of DNA flanked by certain sequences. In wild type Salmonella, the invertible section of DNA is flanked by the *hixL* and *hixR* sequences. The Hin protein binds to these sequences and inverts the section of DNA between them. Haynes *et al.*, instead of using the wild type sequences, used a composite *hixC* sequence. While the *hixL* and *hixR* segments are asymmetrical palindromic, the *hixC* sequences are symmetrical and had a higher binding affinity to the Hin protein and a slower inversion rate. The slower inversion rate was beneficial for their experiment because they were measuring the number of inversions. We will not care about the number of inversions (only the final orientation), but the higher binding affinity to the Hin protein may be useful.

There are also other systems for inverting sections of DNA. Friedland *et al*. used the recombinases Cre and *flpe* in what they called Single Invertase Memory Modules. These modules were flanked by either LoxP sequences or FRT sequences. These worked in similar ways: the Cre protein flipped everything between the LoxP sequences and the *flpe* protein flipped everything between the FRT sequences.

 These systems all do the same thing, though their mechanics are probably different. However, the Hin/*hix* system may be the most beneficial for our purposes because we want to flip, not just single modules, but also several modules together. Though this could possibly be done with the CRE and *flpe* systems, Haynes *et al*. used the Hin/hix system this way. They placed the *hixC* sequence on the outside of and between two different segments of DNA (Figure 3) and the Hin protein was able to flip either one individual segment or both segments.

We will run preliminary experiments to determine which invertase system will work the best for our purposes. We need to determine if we can successfully flip multiple adjacent sections.

Figure 3. Plasmid used by for the Burnt Pancake Problem. (Haynes)

**Inserting the Parts**

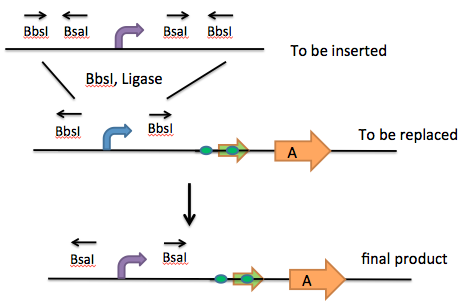
 In order to insert the different parts we will use a system of pairs of Type IIs restriction enzymes (RE) and the Golden Gate Assembly (GGA) method *in vitro*. The part that is to be inserted will have two RE sites facing each other on either side of the part. The outer sites will match sites flanking the part that will be replaced in the gene cassette. When this first RE is added, it will remove the part within the cassette and cut off its recognition sites on the sides of the part being inserted. Then we can add ligase to insert the new part into the cassette. This part can then be removed and replaced with another part using the second RE. Because we are using Type IIs restriction enzymes that cut outside of their recognition site, there are no remaining scars that cannot be recut. Also, we can engineer the overhang made by the cut to be unique so parts will not be inserted into the wrong places.

Figure 4. From “Programming\_Bacteria\_for\_ Optimization\_of\_Ggnetic\_Circuits.pptx. MWSU

For the example in Figure 4, in order to insert the purple promoter in the place of the blue promoter, we would use the restriction enzyme BbsI. Then, because the purple promoter also has the BsaI sites, it can be removed and another promoter can be inserted.

This method can be used for the promoters, RBSs and degradation tags. For the promoter set, each promoter will be flanked by the two RE sites, so that when the RE is added, they will be inserted into the cassettes randomly. The same thing will be done for the RBS set and degradation tag set.

We have found pairs of enzymes that have the same or similar optimal temperatures and work well with the same NEBuffer. These enzymes all cleave to form a four base pair overhang, but if we needed more pairs, smaller overhangs would also work.

**Protocol**

We should be able to put the entire circuit onto one plasmid. With five 26 bp *hixC* sites, four 35 bp promoter sequences, four 88 bp C. dog elements, four 33 bp degradation tags, and four enzymes sequences of 939, 942, 1737, and 909 bp lengths, the pathway will be about 5,300 bp. This is well under the 8 kbp recommended maximum for a plasmid.

Because we cannot look at all the combinations at once due to the vast number of possible combinations we must use a stepwise process, going through each dimension individually.

First, we will insert a promoter into each of the four gene cassettes using the *in vitro* method described above. Then these plasmids will be transformed into the *E. coli*. This *E. coli* will be plated in minimal media, only given L-Arabonolactone (or L-Arabinose) for food, and allowed to grow for several generations. We will then select for the cells that grow the best under these conditions, indicating that their set of promoters is optimal for producing 5-Dehydro-4-deoxy-D-glucarate. We will need to determine a uniform way to select the best growing cells. It may be that the cells with the pathway that leads to optimal growth will have the largest colony, but we will need to determine this once we have growing cells to look at. Then, we will do minipreps, and determine the promoter allele frequencies.

Second, we will repeat this process, inserting the C. dog elements into the four gene cassettes of the plasmids removed for the cells we selected in step 1, select the cells that grow the best and determine the allele frequencies of the C. dog elements. We will then repeat this process for the degradation tags and select the cells that grow best.

For the fourth dimension, the order and orientation of the gene cassettes, we will transform the plasmids from the third step into *E. coli* as well as a plasmid with the Hin gene on it (if we decide to use the Hin/*hix* system). The Hin plasmid will contain a different antibiotic resistance from the plasmid containing the pathway so that we can eliminate the cells that don’t contain both plasmids. We will allow the Hin to express and reorient the pathway. Because we will have *hixC* sites at the beginning and end of the pathway, as well as in between each gene expression cassette, the Hin should be able to change both the order and orientation of each cassette. Then we will do minipreps, separate the pathway plasmids and insert them into *E. coli*. We will culture these cells in minimal media with only L-Arabonolactone for food and select for the cells that grow best.

These finally selected cells should contain the pathways that give the optimal production of 5-Dehydro-4-deoxy-D-glucarate, allowing for the optimal growth. We will do minipreps for these cells, and test for the final allele frequencies as well as for orientation.

**Determining orientation**

In order to determine the final order and orientation of the genes, we will use PCR with specifically assigned primers or barcodes. A unique barcode will be placed at the beginning of each gene expression cassette as well as one at the end. After the shuffling, the barcodes will be in new places. However, when we run PCR with one barcode, the length of the resulting DNA sequences will tell us where it was placed. This can be done with each of the barcodes, so that we can derive the placement and orientation of each gene expression cassette. These barcodes need to be optimally different, but also have similar GC content so that they will all work in the same way. A computer program could be written to determine the best sequences for these barcodes.

**Conclusion**

If this proof of concept works, this method could be used to determine the optimal pathway for a network. Though the particular set up that we find works best for our pathway may not be optimal for another, the method should allow researchers to optimize any pathway for our chosen dimensions, making them more efficient. The increasing complexity of genetic networks has made it more difficult to understand all the interactions between the network and the cell. Lu *et al.* (2009) identified this problem and pointed out the need for a simple platform cell with a minimal genome to test these networks. However, by optimizing the network for fitness within the cell, we do not need to understand all of the interactions between the separate parts and the cell because the cell will determine which set works the best.

**References**

Andersen, Jens Bo, Claus Sternberg, Lars Kongsbak Poulsen, Sara Petersen Bjørn, Michael Givskov, and Søren Molin. 1998. New Unstable Variants of Green Fluorescent Protein for Studies of Transient Gene Expression in Bacteria. *Applied and Enviro. Microbio*. 64:6, 2240-2246.

Friedland, Ari. E, Timothy K Lu, Xiao Wang, David Shi, George Church, James J Collins. 2009. Synthetic Gene Networks That Count. 324, 1199-1202.

Haynes, Karmella A, Marian L Broderick, Adam D Brown, Trevor L Butner, James O Dickson, W Lance Harden, Lane H Heard, Erick L Jessen, Kelly J Malloy, Brad J Ogden, Sabriya Rosemond, Samantha Simpson, Erin Zwack, A Malcolm Campbell, Todd T Eckdahl, Laurie J Heyer and Jeffry L Poet. 2008. Engineering bacteria to solve the Burnt Pancake Problem. *J. Bio. Engineering*. 2:8

Liu, Yang-Yu, Jean-Jacques Slotine, and Albert-Lászaló Barabási. 2011. Controllability of complex networks. *Nature* 473, 167-173.

Lu, Timothy K, Ahmad S Khalil, and James J Collins. 2009. Next-generation synthetic gene networks. *Nature: Biotechnology* 27:12, 1139-1150.

Mogk, Axel, Ronny Schmidt and Bernd Bukau. 2007. The N-end rule pathway for regulated proteolysis: prokaryotic and eukaryotic strategies. *Trends in Cell Bio.* 17:4, 165-172.

Pearson, Brianna, Kin H. Lau, Alicia Allen, James Barron, Robert Cool, Kelly Davis, Will DeLoache, Erin Feeney, Andrew Gordon, John Igo, Aaron Lewis, Kristi Muscalino, Madeline Parra, Pallavi Penumetcha, Victoria G. Rinker, Karlesha Roland, Xiao Zhu, Jeffrey L. Poet, Todd T Eckdahl, Laurie J Heyer and A Malcolm Campbell. 2011. Bacterial Hash Function Using DNA-Based XOR Logic Reveals Unexpected Behavior of the LuxR Promoter. *Interdisciplinary Bio Central.* 3:10, 1-10.