SYNTHETIC BIOLOGY APPROACH TO ENGINEERED COMPETENCY IN *E. COLI*

GENOMICS THESIS

SAMANTHA SIMPSON

Dr. A. Malcolm Campbell, Advisor

Dr. Laurie Heyer, Advisor

Dr. Scott Denham, Advisor

3 April 2009

TABLE OF CONTENTS

LAY AUDIENCE INTRODUCTION

Competency Background

Molecular biology labs across the world rely on the uptake of engineered DNA by bacteria to build and manipulate genetic data that is of importance to basic research. For example, bacteria that take up the insulin gene manufacture insulin for diabetics. Without the ability to manipulate bacteria, scientists would not have been able to manufacture insulin (Gunby 1978) or create a cheap cure for malaria (Keasling 2006). Bacteria that are able to take up engineered DNA are called competent, and there are two types of competency: natural and induced. Natural competency in bacteria refers to the ability to uptake engineered DNA using genetically programmed mechanisms – that is; the cell can switch from a normal growing state to a competent state by itself. Induced competency refers to the ability to uptake engineered DNA through human manipulations, such as electric, heat, or chemical shock. Induced competency can be controlled by researchers but is time-consuming and expensive. Natural competency allows DNA to be taken up by the cell quickly, but is a stochastic process. The act of taking up new DNA is called transformation. Ideally, scientists should be able to have a system of inserting engineered DNA into a competent cell that is reliable, quick, affordable, and controllable. Such a system could be engineered by combining aspects of natural competency and induced competency.

Scientists have thoroughly studied mechanisms of DNA uptake in Gram-positive and Gram-negative naturally competent bacteria. In Gram-positive bacteria, DNA must go through the cell wall and cytoplasmic membrane, and in Gram-negative bacteria, DNA must also go through an outer membrane (Dubnau 1999). In *Bacillus subtilis*, a Gram-positive

bacterium, the first step in transforming new DNA is the binding of double stranded DNA (ds-DNA) to the cell surface, an action that is not dependent on the base sequence of the DNA or on the molecular weight of the DNA. The DNA is then cleaved into pieces that are on average 13500-18000 base pairs long (Dubnau 1972, 1974). The ds-DNA is also converted to single stranded DNA (ss-DNA) before it is transported across the cell wall and cytoplasmic membrane in a linear fashion at the rate of approximately 180 nucleotides/second (see figure 1-1) (Dubnau 1999). Transportation across the outer membrane in naturally competent Gram-negative strands is not thought to change the DNA transport mechanism much because many of the same genes are used in both Gram-positive and Gram-negative naturally competent organisms.

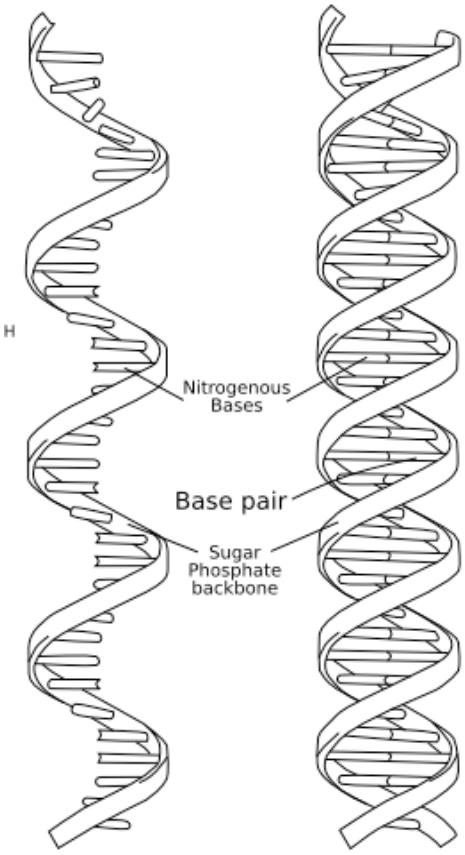

Figure 1-1. Single stranded and double stranded DNA.
http://upload.wikimedia.org/wikipedia/commons/0/04/NA-comparedto-DNA_thymineAndUracilCorrected.png

Natural competency is thought to have evolved to aid the cell in situations of nutritional stress or DNA damage (Johnsborg 2007, Hamoen 2001). DNA taken from the environment can be broken down completely to be used as a carbon source, or broken down into individual nucleotides to replace mutated or destroyed DNA. Cellular competency also is a mechanism by which the cell can acquire new genes and functions from its environment for evolutionary purposes (Johnsborg 2007). Acquiring a new gene would happen when whole genes remained intact when the ds-DNA is chopped into

smaller pieces and made into ss-DNA in preparation to enter the cell. The cell becomes competent in situations where it is nutritionally stressed or senses a high cell density in its surroundings[1], but even these environments are unpredictable. For example, natural competency can turn on when the cells grow on rich media (an environment where all the molecules a cell needs for growth are present) when it is clearly not under nutritional stresses (Van Sinderen 1994 *ComK*). Natural competency has been classified as a stochastic and bistable system[2] (Losick 2008). In *B. subtilis*, competency is controlled by a protein that is found at a concentration of 1 mRNA transcript per cell in competent organisms and 0.3 molecules per cell in noncompetent organisms (Maamar 2007), overall making 15% of *B. subtilis* cells competent in the stationary phase. Researchers have been unable to utilize natural competency's simplistic and fast mechanism to manipulate the introduction of new DNA of interest into microorganisms.

Transformation into an inducibly competent organism is a more controllable method that can be done using electroporation, ploroplasts, and heat shock/$CaCl_2$ treatment in organisms that are not naturally competent. *Escherichia coli* is a commonly used inducibly competent research organism, and a typical procedure for transformation takes 2 days to make the cells competent or between $70-$300 for 1mL to buy competent cells, and another day to transform the cells. The procedure used to make *E. coli* competent produces a low fraction of competent cells (10^{-3} to 10^{-6}) (Hanahan 1985). *E. coli* transformation mechanisms are not well understood, but it is known that Ca^{2+} induced

[1] Having more cells in an area correlates to having more recently dead cells in an area, which often lyse and release DNA into their surroundings. This DNA is available for harvest by competent cells, which is why cells tend to become competent in areas of high cell density.
[2] Bistable means the cells are either competent or not – there is no in-between stage.

competence is mediated chemically and not physiologically (Wenhua 2004). Not much is known about the cellular mechanism of DNA uptake in organisms with inducible competence.

In *B. subtilis*, the research to determine the genes enacting parts of the DNA transport has spanned 18 years. First, the gene ComK was observed as a regulatory gene involved in competency. Protein expression patterns identified ComG, ComC, and ComDE operons as being involved with competency (Dubnau 1991). Later, it was found that ComK expression was dependent on ComL, and that ComK was only produced post-exponential growth (when *B. subtilis* was no longer rapidly dividing) (Van Sinderen 1994 *Molecular*). ComK also appeared to be produced only in high cell density environments and was turned on by ComP and ComA (Grossman 1995). Next, researchers identified SinR, DegU, and ComS as regulating ComK expression (Hamoen 2003), and identified ComS's role in making the competent state last longer (Süel 2006). Not much is known about how the competent state starts, besides that ComL, ComP, ComA, SinR, DegU, and ComS all regulate the production of ComK, which then turns on the production of all genes needed to bind DNA to the cell wall, cleave the DNA, and transport the DNA into the cytoplasm (see figure 2).

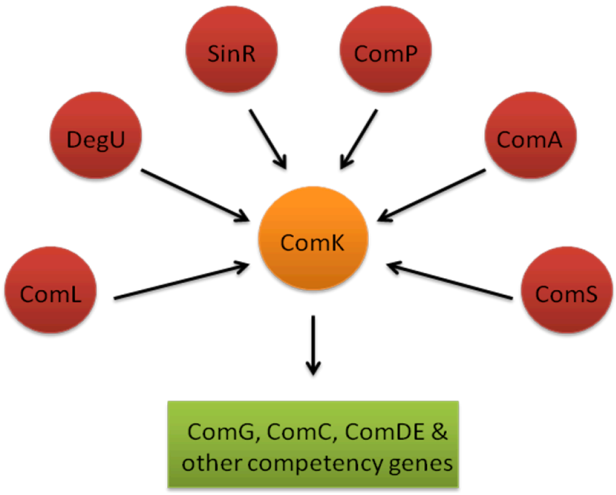

Figure 1-2. Relationship between the genes involved in competency. Arrows represent activation. ComK also activates its own production.

A comprehensive study by Hamoen and colleagues (2002) details how ComK works after it has been activated. Only one copy of ComK mRNA is needed, on average, to induce competency in steady-state *B. subtilis* (Maamar 2007). ComK functions as a transcriptional activator, meaning that after being transcribed into mRNA and translated into a protein, it acts by seeking out 105 genes – some involved in competency, some of unknown function - and increasing the transcription and translation of those genes. ComK accomplishes this monumental task by recognizing and binding to specific sequences of DNA in front of the genes it activates (Hamoen 1998). Attaching to the promoter attracts another protein, RNA polymerase, which begins to make an RNA strand using the DNA as a template. The RNA, after a few slight modifications, then acts as a messenger to other cellular components that make proteins (thus the term mRNA, m for messenger). ComK knows attaches to a specific sequence called a K-box promoter (Hamoen 1998). This K-box promoter sequence appears 1062 times in *B. subtilis*'s genome, when statistically it should only occur about 175 times in a genome that is the same length with the same AT-content of *B. subtilis*'s genome (Hamoen 2002). AT-content refers to the percentage of adenine (A) and thymine (T) bases present in an organism as opposed to the guanine (G) and cytosine (C) DNA bases.

Clearly, there are more K-box promoters than ComK-activated genes. Additionally, not all ComK-activated genes have a K-box promoter in front of them. In fact, only 20% of the 105 ComK-activated genes have a K-box sequence 200 base pairs upstream of the gene, which is the region that generally includes the promoter. Only 30% contain a K-box sequence in the 1000 base pairs upstream of the ComK-activated gene (Hamoen 2002). While this is much lower than expected, it does not take operons into consideration. Operons are strings of genes that are all regulated by one single promoter. For example, a

K-box promoter may be immediately followed by one or more tandem genes that are longer than 1000 base pairs. The first gene may be directly followed by another gene that is controlled by the same K-box promoter, but it would be further than 1000 base pairs away. Making the assumption that genes in the same orientation that do not have transcriptional terminators between them are parts of operons, 45% of ComK-activated genes had a K-box promoter region (Hamoen 2002). 45% of genes are still much lower than expected as ComK is not known to bind to any other DNA sequences (see figure 1-3).

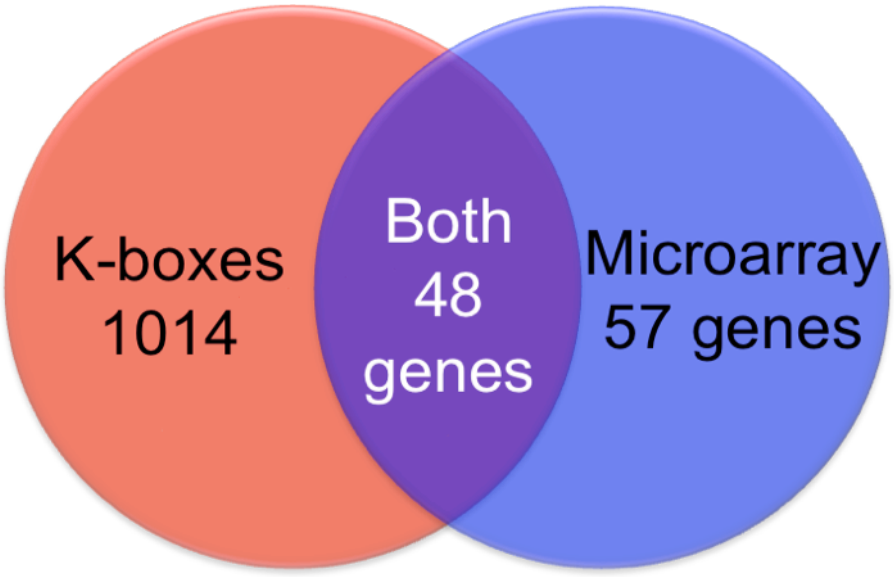

Figure 1-3. Late competency genes are predicted by K-Box sequences in their promoter regions and by ComK activation shown by microarray data. Only 48 genes share these characteristics.

E. coli does not have any known natural competency functions. However, *E. coli* can consume ds-DNA as a carbon source, and *E. coli* has eight genes that are orthologous to late competency genes in other organisms (Palchevskiy 2006). Late competency genes refer to genes activated by ComK that also have known roles in competency. Orthologous genes are those that have similar DNA sequences in two different species and are therefore likely to have similar functions. Another study found that a transcriptional activator commonly used in *E. coli* is used by *Haemophilus influenzae* to initiate competency. This transcriptional activator, the cAMP receptor protein, utilizes one

promoter sequence in *E. coli* and *H. influenzae* for general transcriptional activation, and has a separate promoter sequence in *H. influenzae* for its competency initiation, which indicates that *E. coli* once may have had that competency function as well but it was not beneficial enough for the organism to maintain through evolution (Cameron 2006). Finding that *E. coli* has late competency gene orthologs and an antiquated competency transcriptional activator laid the groundwork for a hypothesis that *E. coli* may contain remnants of a natural competency system.

Math Modeling Background

Scientists can gather numerical data describing aspects of a genetic circuit and use the data to create a hypothetical model of how the cell works. The model can then be used to make predictions and increase understanding about the complexities of the cell. Many cellular processes begin with a slight chemical concentration increase, which in turn leads to the change in conformation of a protein, which is then able to bind to DNA and activate transcription and translation, and these newly produced proteins then go on to have an impact on other aspects of the cell. Determining how strongly a promoting unit reacts to a change of stimulus interests scientists because it allows them to make predictions as to how much end product will be produced. An easy way to determine the strength of a promoter is to measure the production of a protein that is coded by a gene immediately following the promoter. Scientists often measure the production of green fluorescent protein (GFP) indirectly by measuring its fluorescence, which is a simple process done by machines. However, a promoter alone is not enough to ensure that the following DNA-encoded protein gets transcribed and translated. A ribosomal binding site is also needed.

Previous modeling efforts group the promoter and ribosomal binding site (RBS) into one constant (Leveau 2001, Alper 2005), despite the fact that there are a variety of strengths of both promoters and RBSs that could be combined to allow scientists to achieve a variety of transcription and translation strengths.

Previous attempts at modeling competency have focused on all the factors that effect the transcription and translation of ComK in its natural setting. Equations describing ComK leading the cell's transition into the competent phase focus on ComS production while also acknowledging that ComS is not the only contributor to competency (see figure 1-2) (Süel 2006). Research has also looked at changing the natural production level of ComK and ComS. When ComK production is increased by 20 times its normal levels, 100% of *B. subtilis* cells enter competency. When ComS production is increased by 20 times its normal levels, only about 20% of *B. subtilis* cells entered competency, compared to 3% of cells when both ComS and ComK are expressed normally (Süel 2007). Modeling changes in ComK expression and the expression of some activators of ComK show that ComK relies on several activators; however, the easiest way to dramatically increase ComK production is to increase the transcripts of ComK itself instead of increasing one of its activators. No mathematical model has been created to describe ComK's effects on individual late competency genes.

Laboratory Research

Making naturally competent *E. coli* while allowing scientists to retain control over transformation would revolutionize the field of molecular biology. Scientists could take any *E. coli* cell line, make it competent, add engineered DNA, and check that the uptake of DNA

was successful all in a few hours rather than in a few days, and with minimal expense. Naturally competent *E. coli* would also have economic benefits as well, since the only chemical needed to induce competency would be some type of sugar in the growth media.

One feasible way to make *E. coli* naturally competent is to incorporate the *B. subtilis* ComK system into *E. coli*. *E. coli* is a Gram-negative system and *B. subtilis* is a Gram-positive system, but current research indicates that the genetically-regulated competency system should work the same both ways (Dubnau 1999). Incorporating all 105 genes activated by ComK in *B. subtilis* into *E. coli* seemed unreasonable. One can separate these 105 genes into genes that either have a known role in competency or an unknown function (of which there are 79), and into genes that have a known function not involved with competency. Adding 79 new genes into *E. coli* would still be a remarkable accomplishment. Comparing the sequence of those 79 genes from *B. subtilis* to *E. coli* gene sequences yielded 44 orthologs. Of those 44 genes in *E. coli*, the average distance between the gene and the preceding K-box promoter sequence was significantly different from the distance between the K-box promoter sequence and the ComK-activated genes in *B. subtilis*, but large standard deviations in the data caused a large overlap in the confidence intervals of the samples[3]. All 44 of the genes had K-boxes, and were those part of the 48 genes in the middle of the Venn diagram in figure 1-3.

Thirty-five genes activated by ComK in *B. subtilis* did not show significant orthology to any *E. coli* sequences, and those genes may have been involved in competency or not. Deciding which of those 35 genes are necessary for competency would have to be based on

[3] Confidence intervals are reported with a confidence coefficient alpha. Alpha represents the fraction of the time that the corresponding confidence interval will contain the true mean.

tests done after ComK was inserted into *E. coli*. ComK had no orthologs in *E. coli*, so immediately after inserting ComK I performed a test for competency. Then one operon at a time from *B. subtilis* would have to be added to *E. coli*, and more tests for competency would be performed. Once genes required for natural competency in *E. coli* had been determined, those genes and ComK could be integrated into *E. coli*'s genome to make it permanently controllably competent. The project of making *E. coli* naturally, controllably competent seemed feasible.

The first step in the process of making *E. coli* competent is to add the ComK gene from *B. subtilis*. Ideally, a researcher would like to know the amount of ComK produced inside the cell. The best way to guess the amount of a protein produced if the molecules of the protein cannot be physically counted is to determine the strength of the promoter preceding the gene coding for the protein. Luckily, a Registry of Standard Biological Parts developed at MIT contains some basic information about a variety of different promoters and RBSs (Registry 2009). With the knowledge gained from the Registry, a researcher can make an informed decision regarding what promoter-RBS combination he or she wants to use. Because research has found that on average, in competent *B. subtilis* there is one copy of ComK, a weak promoter should be chosen (Maamar 2007). A weak promoter means that putting it in front of the protein of interest only minimally increases the amount of protein produced. To build a construct that has ComK following a promoter and RBS, I selected a plasmid from the Registry of Standard Biological Parts that had an RBS. A plasmid is a circular piece of DNA that a cell can sustain in addition to its own genome. Plasmids have selectable markers and origins of replication. Selectable markers are DNA sequences that code for genes the cell needs to survive, to ensure that all cells growing contain the

plasmid. Origins of replication control how many copies of the plasmid there are in one cell. In the plasmid I used, the selectable marker was ampicillin-resistance, meaning that cells with the plasmid could grow in a medium containing ampicillin. Ampicillin is an antibiotic, so *E. coli* would not normally grow in its presence. The plasmid containing the RBS also had an origin of replication allowing for a high copy number, meaning there were many copies of the plasmid, and thus the ComK gene, in the cell. I digested the plasmid with restriction enzymes, which are proteins that cut DNA strands at particular recognized sites. This made the plasmid one long string of DNA instead of a circular strand of DNA. Then, I added a piece of DNA that coded for the arabinose-inducible promoter that was cut with the same restriction enzymes to the cut plasmid. This process, called a ligation, allows a plasmid to reform that has the promoter and the RBS. The new plasmid was transformed into competent *E. coli*, then purified and cut with restriction enzymes again. This time, I ligated ComK cut with restriction enzymes into the plasmid, and transformed the promoter-RBS-ComK plasmid into *E. coli*.

Just because the *E. coli* is growing on the selectable marker the plasmid contains does not necessarily mean the *E. coli* contains the desired plasmid. Scientists check to see if the plasmid is the desired length in DNA base pairs, as all constructs have known DNA base pair lengths, by running an agarose gel. An agarose gel, with its complex matrix of sugars, when placed in a liquid buffer charged with an electric current allows DNA to pass through it. DNA is slightly negatively charged, so it moves from the negative pole to the positive pole at various speeds depending on how long the DNA is. Longer DNA strands are impeded by the agarose gel's matrix and do not travel as far as shorter strands. A molecular weight marker containing DNA strands of known lengths can be used in one lane of the gel to

check the length of the strands to be analyzed. It appeared that ComK was not ligating behind the promoter-RBS unit, so a ligation of ComK into the empty plasmid (an empty plasmid still contains the selectable marker and the origin of replication) was performed and the length of the insert was checked. ComK is around 610 base pairs, but the length of pieces in the plasmid ranged from 350-450 base pairs, as shown in figure 1-4.

It appeared as though the ComK was excised from the plasmid. This may occur because ComK even without a promoter or RBS in front of it was still being produced enough to exert an influence over the cell – as stated earlier, only 1 copy of ComK mRNA is needed for *B. subtilis* to move into the competent

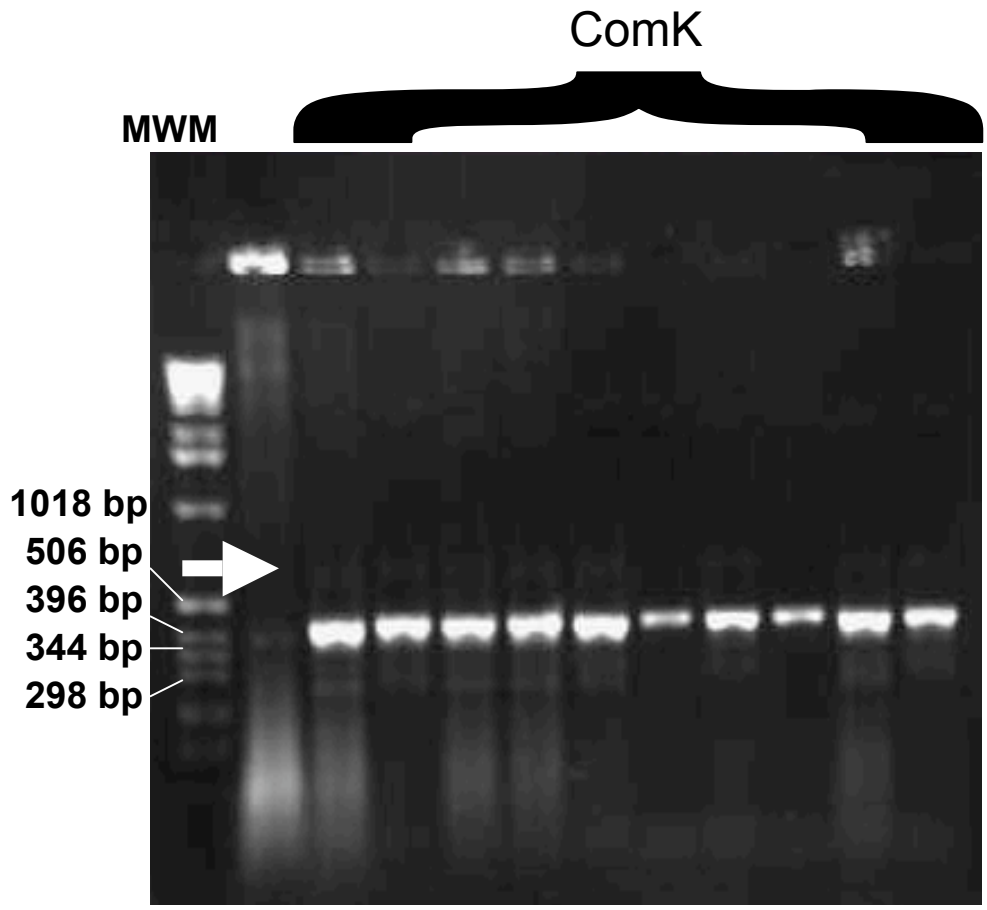

Figure 1-4. A picture of the agarose gel with the plasmids from 10 different E. coli colonies containing the putative ComK insert. The molecular weight marker is in the lane labeled MWM, and the known base pair lengths are to the left of the image. The white arrow denotes where ComK would be expected to appear.

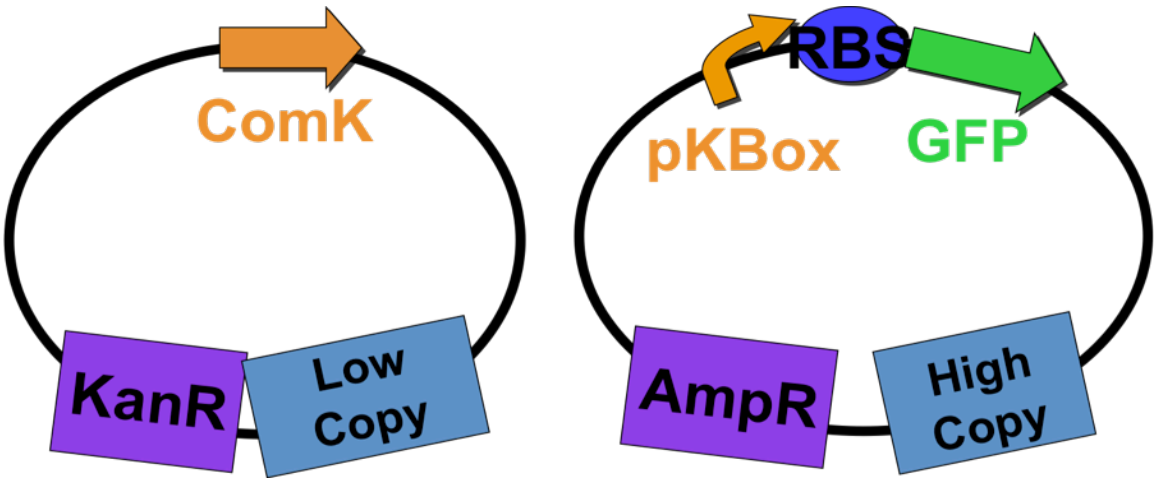

Figure 1-5. The two-plasmid scheme to test ComK activation of the pKBox. The plasmid on the left has the ComK gene, is on a low copy number plasmid with kanamycin resistance, while the right plasmid has the pKBox promoter followed by GFP on an ampicillin resistant plasmid and a high copy number plasmid. GFP can be measured to test the induction of the pKBox promoter.

state. Because the competent state requires that the cell be in stationary phase (not dividing), it is possible that when produced in *E. coli*, ComK switches the cell to the stationary phase as a precursor to its achieving a competent state. Cells in the stationary phase would not produce colonies. Therefore, any growing colonies would be the result of an excision of part of the ComK gene. While frustrating, finding that ComK was excised in a high copy number plasmid was promising as it indicated that ComK may have an effect on *E. coli*'s cellular processes.

Because high copy number plasmids resulted in excised ComK or in cells that did not grow, ComK had to be inserted into a low copy number plasmid. Using an interchangeable plasmid designed to be compatible with the Registry of Standard Biological Parts (Shetty 2008), a low-copy number plasmid with kanamycin-resistance (a different antibiotic) was created, and ComK was successfully inserted. See figure 1-5 for an

illustration of the two plasmid scheme. The process of building the plasmid and inserting ComK took over three months!

To compare the strength of the interaction between ComK and the K-box promoter region to other promoters, a plasmid containing the K-box promoter sequence followed by the GFP gene needed to be co-transformed into *E. coli* with the newly designed ComK promoter. A co-transformation means that two different plasmids are inserted into competent cells at the same time. To ensure that both plasmids are taken up and maintained by the cell, each plasmid in a co-transformation has a different selectable marker and origin of replication. In this example, the plasmid containing ComK had a kanamycin-resistant selectable marker and the plasmid with the K-box promoter sequence had an ampicillin-resistant selectable marker. After the co-transformation, the cells were grown in the presence of ampicillin and kanamycin to ensure that both plasmids were taken up.

To measure the strength of the interaction, fluorescence of the GFP and optical density measurements were taken of cells with both the ComK and K-box promoter plasmids. In order to compare the level of fluorescence with other promoter constructs already available in the Registry of Standard Biological Parts, fluorescence and optical density data was taken for 21 promoter-RBS combinations followed by GFP. The results are

shown in figure 1-6 below.

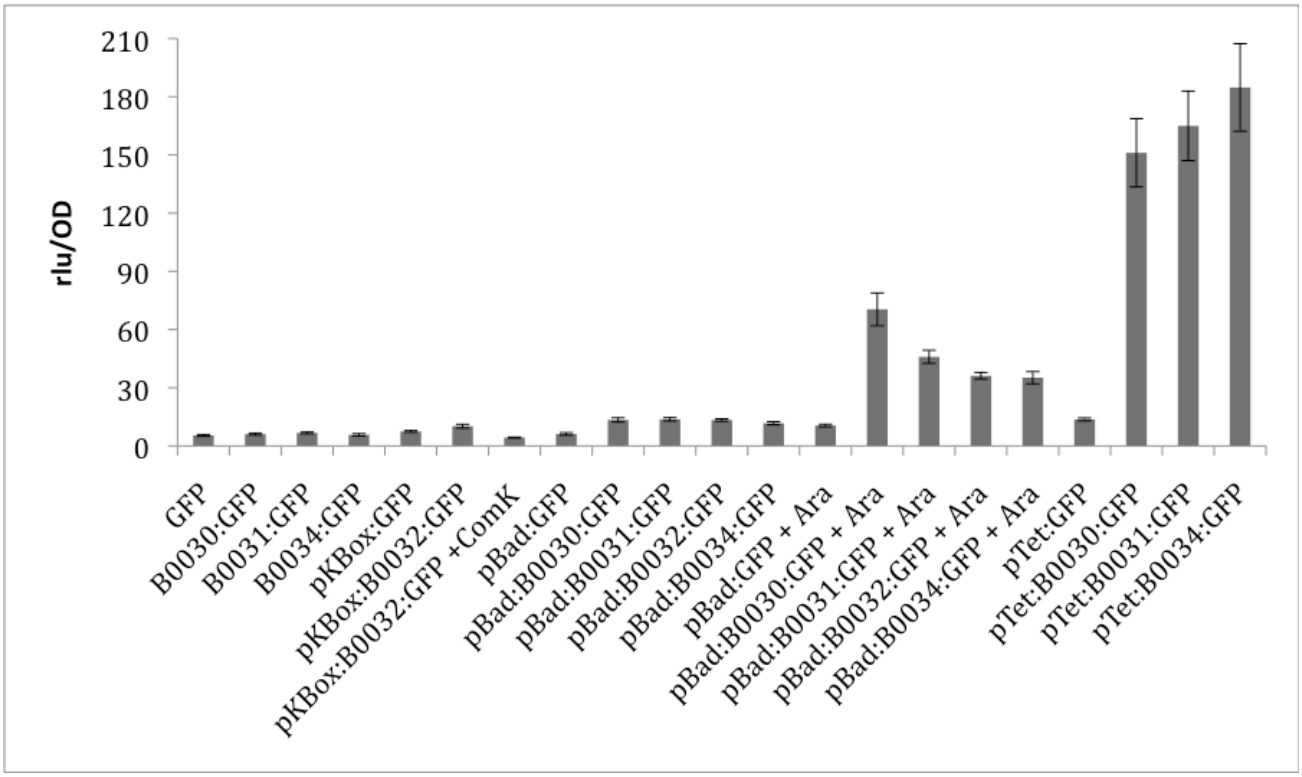

Figure 1-6. Fluorescence / optical density data for 21 promoter-RBS constructs, including the K-box promoter (ComK-Inducible) with and without induction from ComK. Bars are one standard error.

I determined where statistically significant differences were by analyzing this data. Statistical significance is defined as a p-value of 0.01 or lower when performing a T-Test between two data points taken on repeat tests. The baseline GFP fluorescence is represented by the None:None combination. The only RBS by itself that evidenced significantly more fluorescence than the baseline fluorescence was B0031. No None:RBS comparisons supported the hypothesis that one RBS without a promoter was significantly stronger than any others. The RBS that results in the strongest fluorescence when used in conjunction with one promoter is not necessarily the strongest when used with a different

promoter, as would be expected. The RBS data does not point to one RBS as clearly being "the strongest".

Once promoters were taken into consideration, however, differences emerged. The strongest promoter was tetracycline-repressible, then the arabinose-inducible promoter with arabinose, then the arabinose-inducible promoter without arabinose, then the ComK-inducible promoter, and finally the ComK-inducible promoter with ComK. These data indicate that when the ComK-inducible promoter borrowed from *B. subtilis* is inserted into *E. coli*, ComK no longer acts as an activator. This may be because the *B. subtilis* system is more complex than previously thought and more than ComK is needed to create a transcriptional activator.

ComK may be having an effect on *E. coli*'s cellular processes, but tests to see if inserting ComK in *E.coli* causes competency have not shown that competency is that effect. With limited time, I was able to perform three separate competency tests. There are currently no standard procedures to test competency, and there are many variations to the procedures that I performed, so my results by no means allow one to reach a conclusion that *E. coli* with the ComK gene inserted is not competent. More competency tests should be done, as well as inserting a promoter in front of ComK to make production of the ComK protein controllable.

Additional questions arise about RBSs and promoter strengths. The Registry is a service that is maintained by the scientists who use it, but its maintenance is not standardized. For example, all the RBS web pages on the Registry state the RBS efficiencies relative to B0034, which is listed as having efficiency 1.0. My data on RBS strengths, in addition to having no correlation within my experiments, had no correlation with the data

listed on the Registry (see table 1-1). In fact, the RBS the Registry maintains is the strongest (B0034 with efficiency 1.0) was the weakest of all RBSs tested when paired with the arabinose-inducible promoter, the arabinose-inducible promoter with arabinose, and when testing RBSs with no promoters. The RBS the Registry maintains is the weakest (B0031 with efficiency 0.07) was the strongest RBS I tested by itself, and the second strongest RBS when in combinations with the arabinose-inducible and arabinose-inducible with arabinose promoters. Further inspection of the RBS data reveals that RBS strengths were calculated in conjunction with one promoter that is activated by the *cI* protein (Weiss 2001). As my research has shown, different RBS's appear to have varying strengths when combined with different promoters, and this should be made apparent on the Registry.

Just as confusing is the lack of information about promoters to be found on the Registry. The only promoter I tested with any information was the tetracycline-repressible promoter, which was listed as having a medium strength. Once again, more numerical data could benefit the site, especially because the promoters seem to have the most effect on the production of the protein of interest.

My senior thesis leaves a lot of competency-related questions for future genomics majors to explore. First, what is the mechanism of ComK's apparent inhibition of the K-box sequence promoter? Tests need to be done to show that ComK mRNA is produced, that the mRNA is translated into protein, and that the ComK protein binds to the K-box sequence. If ComK actually does act as a repressor of the K-box sequence promoter, then more research needs to be done into the complexity of the *B. subtilis* system. Another unknown molecule may be used in *B. subtilis* that, when combined with ComK, acts as an activator of genes needed for competency. Alternatively, a protein present in *E. coli* that is not in *B. subtilis*

may negate ComK's activities in the cell, thus making ComK appear to act as a repressor. Additional tests for competency need to take place, including adding various other genes from *B. subtilis* to *E. coli* to see if a competent system can be created.

INSERTING A COMPETENCY GENE INTO *E. COLI*

Abstract

Preparing *E. coli* to become competent to uptake new DNA is an expensive and time-consuming process. Chemically competent *E. coli* must be coaxed with extreme temperature fluctuations and chemicals to uptake new DNA. *Bacillus subtilis* can become naturally competent via a transcriptional activator gene called ComK that activates genes preceded by a K-box promoter. Of 79 published competency genes activated by ComK, 44 have orthologs (e-value < 0.01) in *E. coli.* While the distances between the genes and the K-box promoters in *E. coli* are significantly different ($p = 0.015$) from the distances between the genes and the K-box promoters in *B. subtilis*, the average of the set of data points for both lies in the other's 80% confidence interval. Insertion of ComK on a high copy number plasmid into *E. coli* resulted in no cell growth and indicated that the gene may force *E. coli* into stationary phase, which is how ComK works in *B. subtilis*. I measured GFP fluorescence of cells that were co-transformed with ComK and a plasmid containing the K-box promoter followed by GFP. Finally, to see if the addition of ComK in *E. coli* induced competency, I included a plasmid containing RFP in the liquid growth media and observed if any red *E. coli* colonies grew. This work may permit scientists to induce competency in *E. coli* with the addition of IPTG or another transcriptional activating molecule.

Introduction

Natural competency in *B. subtilis* is a well-studied mechanism. The completely genetically controlled change into the competent state is contingent on the concentration of

the competency transcriptional activator called ComK rising above some threshold level within the cell (Hamoen 1998, 2002). Research has shown that in non-competent *B. subtilis*, ComK has an average of 0.3 mRNA transcripts per cell, and in competent cells, ComK has an average of 1.0 mRNA transcripts per cell (Maamar 2007). In general, competency occurs in 3.6 ± 0.7% of all *B. subtilis* cells (Süel 2006), and this is a stochastic process (Losick 2008). ComK has several activators which, when amplified, increase the likelihood of achieving competency (Hamoen 2003). One such activator, ComS, not only increases the likelihood of competence, but it also increases the length of time a *B. subtilis* cell remains competent (Süel 2006). However, the best way to increase competence is simply to increase the amount of ComK coding sequences in the cell – by increasing ComK transcripts by 20 times the original amount, 100% of cells entered competence (Süel 2007).

ComK activates 105 genes in *B. subtilis*, 79 of which could be involved in competency (Hamoen 1998). ComK functions as a tetramer that recognizes the motif $A_4N_5T_4N_xA_4N_5T_4$ where X can either be 8, 18, or 31 (Hamoen 2002). The $A_4N_5T_4$ sequences at known as AT-boxes and the sequences recognized by ComK are known as K-boxes, and there 1062 K-boxes in *B. subtilis*'s genome, which is 6 times the expected number of K-boxes given B-subtilis's genome length and AT-content (Hamoen 1998). However, having a K-box in the promoter region is not necessarily a good indicator that a gene is activated by ComK. Only 8% of genes with K-boxes in their putative promoter regions in *B. subtilis* are activated by ComK, and of the genes activated by ComK, only 45% of genes in operons have a K-box within 200 bps upstream of where the operon starts (Hamoen 1998).

Natural competency is thought to have evolved in order for the cell to respond to nutritional stress (using DNA as a carbon source), for evolutionary purposes (Johnsborg

2007), or for DNA damage repair (Hamoen 2001). In order for an organism to maintain cellular mechanisms for natural competency, the advantage imparted by competence must override energy expenditures necessary to switch the cell to the competent state (Losick 2008). *E. coli* is not a naturally competent organism, yet it contains some vestiges of a naturally competent system. *E. coli* has orthologs to 8 competency genes, 6 of which are activated by ComK in *B. subtilis*. Wild type *E. coli* grown on minimal media containing only extracellular DNA as a carbon source grows 40 fold more than *E. coli* with mutated versions of the competency gene homologs, supporting the idea that *E. coli* may be able to uptake DNA for nutritional purposes (Palchevskiy 2006). Additionally, *E. coli* has an ortholog to a competency transcriptional activator in *H. influenzae*, called the cAMP receptor protein (CRP). In *E. coli* CRP recognizes a different promoter sequence than in *H. influenzae*, which indicates that its role as a competency transcriptional activator has been lost (Cameron 2006). Despite *E. coli*'s apparent evolution away from natural competency, the systems that are used in Gram-negative bacteria such as *E. coli* for natural competency are similar to those used by the Gram-positive bacteria *B. subtilis* (Dubnau 1972, 1974, 1999). Overall, natural competency in bacteria is widespread and highly conserved (Dubnau 1999). While *E. coli* is not known to be naturally competent, indications that it evolved from a naturally competent organism are present and deserve more exploration.

Materials and Methods

Genomic DNA Isolation

I used Zymo's YeaStar Genomic DNA Kit (Cat. No. D2002) to isolate fresh *B. subtilis* genomic DNA from a TSA plate. I followed protocol II, except I allowed the digestion process to

continue for 4 hours, and at step five I switched to the Campbell Lab ethanol precipitation of DNA protocol:

http://www.bio.davidson.edu/courses/Molbio/Protocols/clean_short.html

Plasmid DNA Isolation

I used Promega's Wizard Plus SV Minipreps (Cat. No. A1460) to isolate plasmid DNA from overnight cultures.

Digestions

I used the Campbell Lab digestion protocol, allowing digestions to run from 2 hours to 14 hours, and often exceeding the recommended volume of 20μL:

http://www.bio.davidson.edu/courses/Molbio/Protocols/digestion.html.

Ligations

I used the Campbell Lab ligation protocol using Promega's 2X ligation buffer (Cat. No. C6711) and often exceeding the recommended volume of 10μL when insert DNA was too dilute. I never exceeded 20 μL of ligation. :

http://www.bio.davidson.edu/courses/Molbio/Protocols/ligation.html

Transformations

I used the Campbell Lab heat shock transformation protocol, using JM109 cells from Promega (Cat. No. L2001):

http://www.bio.davidson.edu/courses/Molbio/Protocols/transformation.html

Polymerase Chain Reaction (PCR)

I used the Campbell Lab PCR instructions using Promega's GoTaq Green Master Mix (Cat. No. M7122): http://www.bio.davidson.edu/courses/Molbio/Protocols/pcr.html#monster.

All primers were ordered from MWG Biotech in 100pM concentrations.

Gel Electrophoresis

I used 0.5x TBE buffer to run gels using Invitrogen's 1 Kb DNA ladder (Cat. No. 15615-016).

Gel Elutions

To perform gel elutions, I used QIAquick Gel Extraction Kit (Cat. No. 28704) and followed Qiagen's protocol.

Determining DNA Concentration

To determine the DNA concentration of a sample, I used the NanoDrop ND-1000 spectrophotometer and its provided software.

Fluorescence and Optical Density Readings

Fluorescence data was taken using an FLx800 microplate fluorescence reader, measuring at an excitation of 485/20 and an emission of 528/20. Optical density was taken using an ELx808 at the 595 endpoint. Arabinose was added to arabinose-inducible populations at a concentration of 0.002% w/v.

Results

I wrote a Perl computer program to search through the *B. subtilis* and *E. coli* genomes and identify all K-box motifs ($A_4N_5T_4N_xA_4N_5T_4$ where X can either be 8, 18, or 31) (see appendix 1). According to the Hamoen paper (2002), there are 1062 K-boxes in *B. subtilis*'s 4.1 Mb genome, using the sequence from GenBank at NCBI (2008). I found a 4,214,530 bp sequence with 1078 K-boxes for *B. subtilis* using GenBank and the K-box finding algorithm. My program found one more match (13/16 A's and T's present) to the small (8 base pairs between the two AT-boxes) K-box formula, one more match (14/16 A's and T's) to the large (31 base pairs between the two AT-boxes) K-box formula, and 14 more

matches (13/16 A's and T's) to the large K-box formula. I used the same program to find K-boxes in *E. coli* K-12 substr. DH10B, which has 4,686,137 bps and found 88 small K-boxes, 88 medium K-boxes, and 105 large K-boxes, or 281 total.

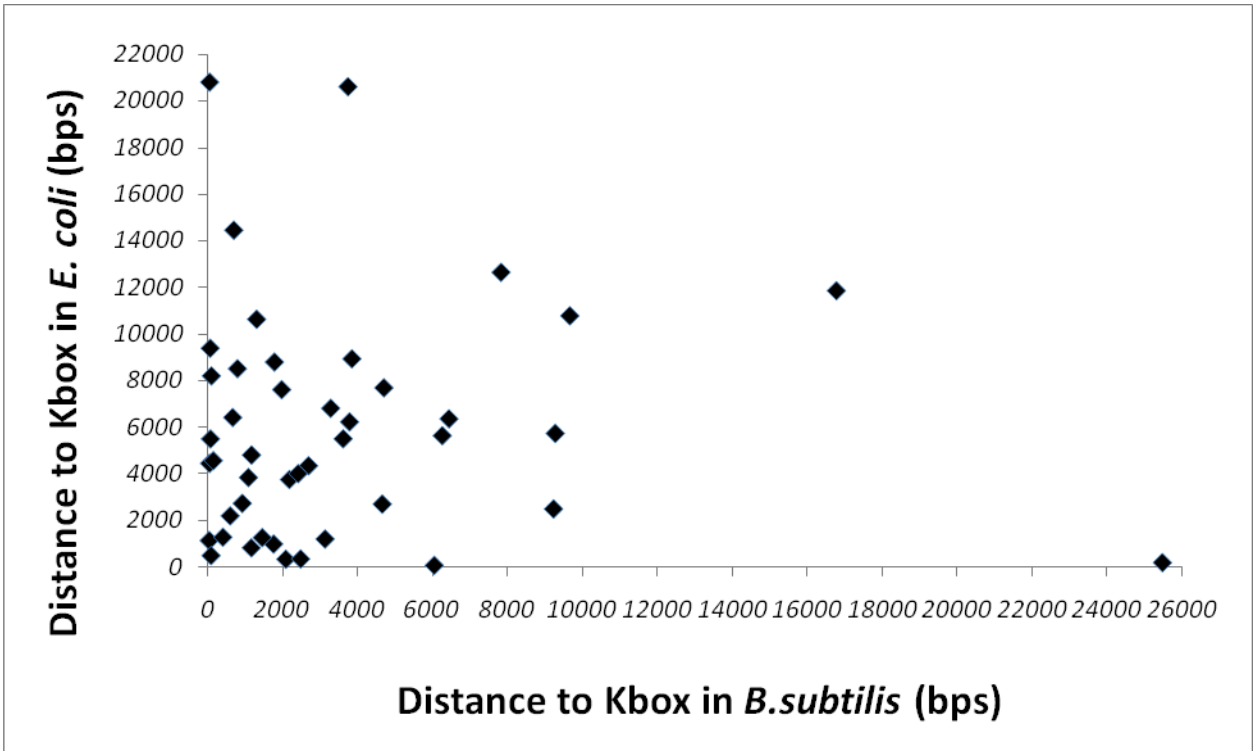

Figure 2-1. Information from the Hamoen paper (2002), my Perl K-box finder (appendix 1), and NCBI BLAST were analyzed to create this chart. Each dot represents one of the 44 ComK-activated genes that had an ortholog in *E. coli* with a BLAST e-value of less than 0.01. The distance from the nearest K-Box to the gene in *B-subtilis* is plotted against the distance in *E. coli*. The data is largely scattered, indicating an overlap in the 80% confidence interval of the average K-Box to gene distance in *E. coli* and *B. subtilis*.

I then used NCBI's BLAST to find sequences for the 79 genes activated by ComK in *B. subtilis* and performed a protein BLAST with *E. coli*'s known proteins to find any genes that had orthologs with an e-value < 0.01. 44 genes satisfied this requirement, and I determined how far away the closest upstream K-box was to those genes. A 2-sample Z Test of those 44 genes and their orthologs in *B. subtilis* revealed that the distances between the gene and the K-box, when compared, had p = 0.015. The 80% confidence interval for the average *B.*

subtilis K-box to ComK-activated gene length is 3566 ± 6140 base pairs. The 80% confidence interval for the average *E. coli* K-box to gene ortholog length is 5824 ± 6331.

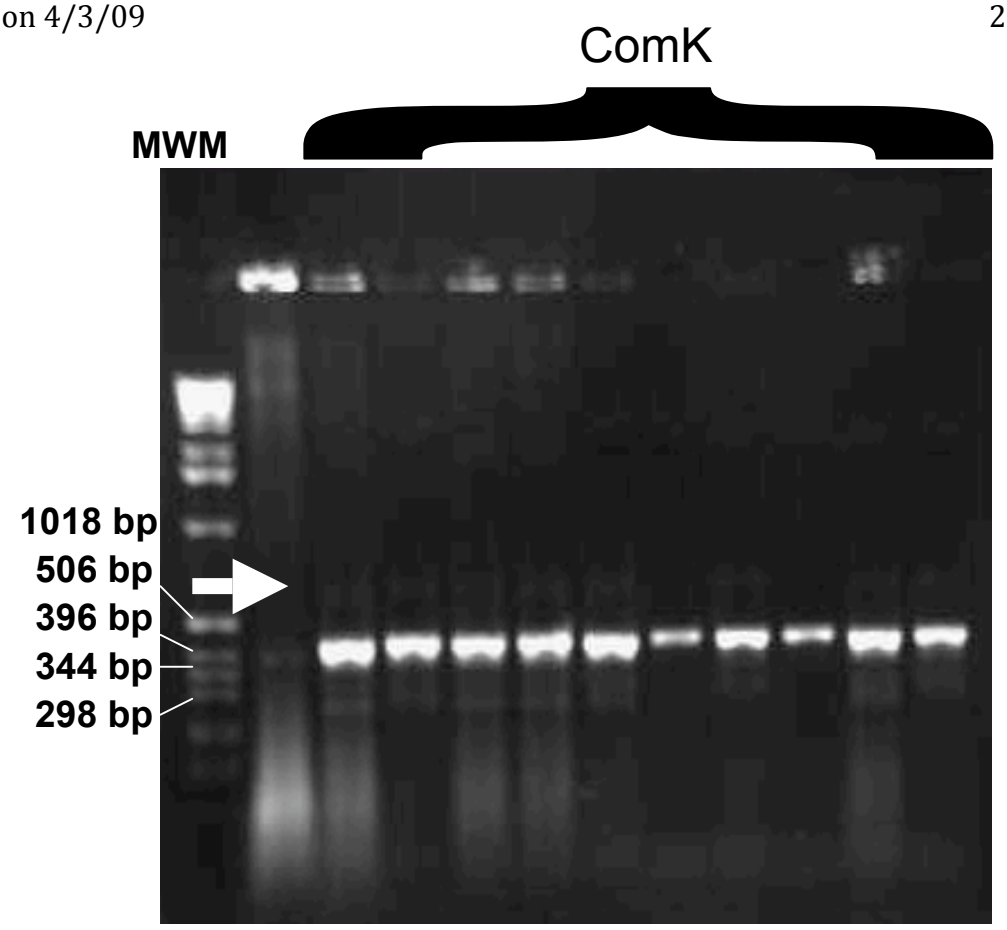

Figure 2-2. Picture of the agarose gel with plasmids from 10 different *E. coli* colonies containing the putative ComK (580bps, marked by white arrow). The first lane contains a 1Kb MWM.

Information about ComK-activated gene orthologs and distances to the K-box sequence is presented in figure 2-1.

I used PCR to obtain ComK, a promoter sequence containing K-boxes referred to as pKBox, a Kanamycin resistance gene, and a low-copy origin of replication. Primers and DNA sources used are given in table 2-1 below.

Table 2-1. Promoters and templates used to clone various DNA sequences flanked by BioBrick ends (Knight). Where applicable, the start of the gene or promoter is in all caps. *Data taken from the Registry of Standard Biological Parts.

Gene	Forward 5'-3'	Reverse 5'-3'	Template
pSC101 Origin of Replication (I50042*)	GCATgctagcctgtcagaccaagtttacg agctcgc	GCATgctagcaacacccttgtattactgt ttatg	pSB4A5*
Kanamycin Resistance	GCATgaattcgcggccgcttctagac tgatccttcaactcagc	GCATctgcagcggccgctactagtatt attagaaaaactcatcgagc	pSB1AK3*

ComK	GCATgaattcgcggccgcttctagAT GAGTCAGAAAACAGACGC	GCATctgcagcggccgcaactagta CTAATACCGTTCCCCGAGC	*B. subtilis* genome
pKBox	GCATgaattcgcggccgcttctagaC GTTGTGCTCAATTTTTTC	GCATctgcagcggccgctactagtaC AGATATAACAGAGACGAAC	*B. subtilis* genome

The pKBox promoter was designed from the ComE operon in *B. subtilis*. It begins 10 base pairs upstream of the actual K-box sequence, continues through the Shine-Dalgarno sequence, and ends at the start codon of ComEC. ComK has two internal EcoR1 sites, and pKBox has one internal Xba site.

ComK, a 580bp part, was initially cloned into pSB1A2, a high copy number plasmid with a pMB1 origin of replication. After isolating the plasmid DNA and digesting with restriction enzymes meant to excise the plasmid, the gel electrophoresis results were as found in figure 2-2. Part of ComK appears to be excised. The plasmid was empty before the attempted ligation of ComK PCR product. The PCR product itself was the expected 580 base pair length. Several cellular mechanisms can lead to the excision of DNA, one of the most common being inverted repeats in the DNA self-excising (Reddy 200). Figure 2-3 is a dot plot of ComK's sequence against itself to illustrate where inverted repeats are.

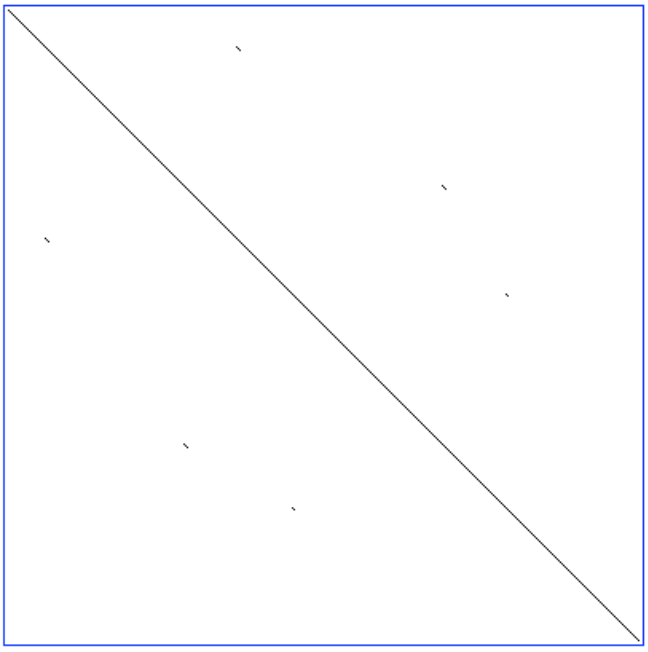

Figure 2-3. A dot plot comparing the ComK DNA sequence to itself. This shows that there are no large sequence inversions that would self-excise.

After building a low copy number, kanamycin-resistant plasmid (Shetty 2008) ComK was successfully cloned. To build the plasmid, I PCR'd the BioBrick base vector (which has ampicillin resistance flanked by two Nhe1 sites, and no origin of replication) from the Spe-Pst BioBrick end all the way to the EcoR1-Xba BioBrick end. I then digested the PCR product with EcoR1 and Pst and inserted the low-copy number origin of replication into the BioBrick insert site. In an empty pSB1A2 plasmid, I ligated together the kanamycin-resistance gene and the low-copy number origin of replication. I then PCR'd this fragment so it would be flanked by Nhe1 sites. I cut the PCR product and the BioBrick base vector with the low copy number origin of replication insertion (which removed the ampicillin resistance gene) with Nhe1, and inserted the kanamycin-resistance gene-low copy origin PCR product. I then cut the engineered BioBrick base vector with Xba and Spe and performed a shrimp alkaline phosphorylation to be sure the cut ends of the vector did not ligate together. I then cut ComK with Xba and Spe and inserted it into the BioBrick insert site to make my completely engineered plasmid.

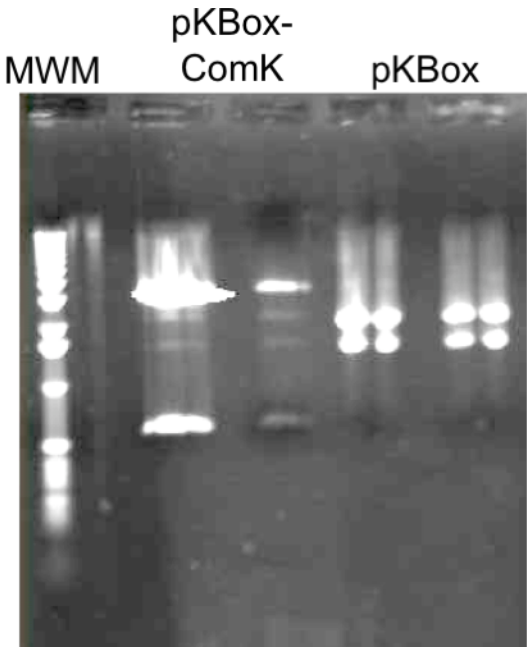

Figure 2-4. Co-transformation of ComK and pKBox plasmids. The first lane contains a 1Kb MWM. The second and third lanes contains the co-transformation, and the last two lanes contains pKBox. All samples were digested with Xba and Spe.

The low copy number, kanamycin-resistant ComK plasmid was then co-transformed with a pSB1A2 plasmid containing pKBox : B0032 (Registry 2009) : GFP (see

figure 2-4). *E. coli* containing the pKBox plasmid without ComK grew at the same time and was tested for fluorescence and optical density. Results are shown in table 2-2 below.

Table 2-2. Average Fluorescence and OD data for pKBox ± ComK *E. coli* samples

Plasmid	# Trials	Fluorescence (rlu)	Optical Density (OD)	rlu OD^{-1}
pKBox:B0032:GFP + ComK	12	4.3	0.999	4.29
pKBox:B0032:GFP	12	8.1	0.856	10.18

After fluorescence and OD data collection, samples were tested to ensure *E. coli* had maintained both the ComK and the pKBox plasmid.

I tried three separate methods of my own design to test for competency. Previously, researchers had tested ComK production and concluded that the cell was competent because of the amount of ComK mRNA, but a simple test checking the uptake of DNA was needed. First, I put the ComK plasmid in *E. coli* in LB media with ampicillin, kanamycin, and 450ng of a pSB1A2 plasmid containing the red fluorescent protein (RFP) gene. Cells grew, but none contained the pSB1A2 plasmid.

Second, I grew up *E. coli* containing the ComK plasmid overnight in LB media with kanamycin. I pelleted the cells and resuspended them in 100µL water, added 450ng of a pSB1A2 plasmid containing RFP, and allowed this mixture to sit for 15 minutes before plating the solution on two plates: one with just kanamycin, and one with kanamycin and ampicillin. Lawns of bacteria grew on the kanamycin plates, but nothing grew on the ampicillin plates.

Third, I grew up *E. coli* containing the ComK plasmid for four hours in LB media with kanamycin. I pelleted the cells, resuspended in water, performed a 42˚ 45 second heat

shock, and waited 15 minutes before plating the solution as described above. Lawns of bacteria grew on the kanamycin plates, but nothing grew on the ampicillin plates.

Discussion

ComK's inability to clone into a pSB1A2 plasmid may indicate that the cell is unable to grow with large amounts of ComK present. In *B. subtilis*, heightened production of ComK (from 0.3 mRNA molecules/cell to 1 molecule/cell) leads the cell into stationary phase (Maamar 2007). All colonies selected for inspection via gel electrophoresis had inserts of approximately 350-450 base pairs, which may indicate that the colonies that grew excised part of ComK, which normally has 580 base pairs. The similarly sized excision could be due to *E. coli* recognizing a part of ComK's sequence, or a pattern in the ComK gene that would lead to its being excised at a particular spot every time. *E. coli* may retain or may have adapted some sort of cellular recognition of ComK that would lead it into stationary phase.

The fluorescence and optical density data also indicate that ComK has some sort of effect at the cellular level, although not the expected one. In *B. subtilis*, ComK is widely known as an activator of transcription that binds to a sequence called a K-box (Hamoen 1998), although less than 50% of late competency genes have the K-box sequence within 1,000 base pairs of their start codon (Hamoen 2002). The K-box sequence used in pKBox in this research is approximately 470 base pairs away from where the start codon of GFP is located, which is closer than the average distance from the K-box to all ComK-activated genes, which is 3,028 base pairs (Hamoen 2002). Despite this, the apparent function of the pKBox-ComK promoter system reversed when not in *B. subtilis*. This, along with the Hamoen findings about the unreliability of using K-boxes to predict the location of late

competency genes (Hamoen 2002), prompts a further look into *B. subtilis*'s competency mechanisms. Perhaps another protein is involved in *B. subtilis* that, when combined with ComK, is actually responsible for the increase in cellular competency. As mentioned before, *E. coli* may have a cellular response to ComK, which would correlate with the finding that the ComK DNA was excised from the plasmid.

Another explanation for the lower fluorescence of the co-transformed cells can be found in figure 2-4. The brightness of the bands indicate that there is less pKBox promoter plasmid than ComK plasmid, which is the opposite of what is expected considering that ComK is on a low copy number plasmid and pKBox is on a high copy number plasmid. A lower amount of plasmids containing GFP following the pKBox promoter may be why the GFP fluorescence is significantly lower in the co-transformed cells than in cells that contain only the pKBox promoter plasmid. There is no direct comparison between cells containing only GFP and the co-transformed cells, so it is unclear whether the low amounts of pKBox-GFP plasmid contributed to the fluorescence of co-transformed cells being significantly lower than the fluorescence of plain GFP.

The hypothesis that ComK plays a role as a transcriptional repressor is strengthened by the missing DNA when ComK was in a high copy number plasmid. It is unusual for an event like DNA excision to occur unless it was a directed excision, favorable to the cell's growth and stability. However, DNA does not effect a cell's growth – protein does. The fact that the DNA was excised supports the idea that ComK is being produced as a protein by *E. coli* and is having some effect, although further tests would be required to prove its presence, such as an assay for ComK mRNA and protein.

While the evidence strongly supports a role for ComK in *E. coli*, the role is not the expected role of a transcriptional activator moving towards competency. The competency assay was negative, and ComK does not activate transcription of GFP when placed under control of a promoter with a K-box. However, these experiments with ComK in *E. coli* may allow *B. subtilis* researchers insight into the mechanisms of natural competency for *B. subtilis*. Future research in this field may look for proteins that interact with ComK, or other DNA sequences with which ComK might bind.

MODELING PROMOTERS AND RBSs IN *E. COLI*

Abstract

In the past, the strength of various biological promoter-RBS combinations has been modeled as a unit. Separating the strength of the promoter and the RBS would allow biologists to more specifically control the transcription of DNA sequences of interest. I constructed 22 plasmids with various promoters and RBSs followed by GFP and measured fluorescence to compare the strength and transcriptional control of various combinations. Statistical analyses were performed to determine the most importance differences. I found that various RBSs, once present, do not influence transcription levels consistently. Therefore, promoters by themselves are better indicators of transcriptional control and choice of RBS appears to be irrelevant. A novel promoter-activator system borrowed from *B. subtilis* was also tested. In *E. coli*, the introduction of a plasmid containing the inducer molecule appears to repress promoter activity ($p < 0.01$).

Introduction

Biologists often want to express a protein at a certain known and controllable level. Expression of proteins can be controlled by varying both the promoters and ribosomal binding sites (RBSs) preceding the gene coding the protein of interest. Several characterized promoters and RBSs are available for molecular biologists from the Registry of Standard Biological Parts.

Previously, modeling efforts to inform scientists of the strength of the promoting unit used have grouped the promoter and RBS into one constant. One method allows

scientists to calculate the strength of the promoter using fluorescence and optical density data taken when the green fluorescent protein (GFP) follows the promoter and RBS of interest. The biologist can calculate the strength of the promoting unit (P) using the following equation:

$$P = f_{ss} \cdot \mu \cdot (1 + \mu/m)$$

Where f_{ss} is the fluorescence in relative light units (RLU) divided by optical density (OD), μ is the growth rate of the bacteria per hour, and m is the maturation constant of GFP (because some GFP variants do not fluoresce immediately after the protein is formed) (Leveau 2001). Experimentation has found m to be 1.5 h^{-1} in non-engineered GFP (Andersen 1998). In the past, scientists have used nucleotide analog mutagenesis to create promoting units of varying strength for both bacterial and eukaryotic systems (Alper 2005). However, this method may not be extremely useful for scientists interested in using the Registry of Standard Biological Parts, because promoters and RBSs are separated into two different categories which researchers must choose.

The Registry is a repository for standardized biological parts. Each part in the Registry is flanked by a BioBrick prefix and suffix that consists of standard restriction enzyme sites that can easily be digested and ligated with other registry parts (Andersen 1998), as shown below:

5' --gca GAATTC GCGGCCGC T TCTAGA G --insert-- T ACTAGT A GCGGCCG CTGCAG gct--

--cgt CTTAAG CGCCGGCG A ACATCT C --insert-- A TGATCA T CGCCGGC GACGTC cga--

EcoRI NotI XbaI SpeI NotI PstI

Currently, the RBSs have relative strengths posted on their sites. The RBSs and their relative strengths as posted on the Registry are listed in table 3-1 below:

Table 3-1. RBSs and RBS efficiencies as given by the Registry for Standard Biological Parts. Efficiencies are based on the efficiency of B0034, which is the RBS used in the Elowitz repressilator (2000).

RBS	Relative Strength (to B0034)
B0030	0.6
B0031	0.07
B0032	0.3
B0033	0.01
B0034	1.0

The RBS efficiency values come with no references or other data, which make the Registry system seem unreliable. Some promoters have information available. Table 3-2 below lists common promoters and information given on the Registry.

Table 3-2. Promoters and promoter strengths as given by the Registry for Standard Biological Parts. Some information is missing.

Promoter	Relative Strength
R0040 – TetR Repressible	Medium (Lutz 1997)
I0500 – Arabinose Inducible, AraC Repressible	Weak-Medium w/ 0.2% arabinose (Johnson 1995)
I13458 – Arabinose Inducible	?
I13453 – Arabinose Inducible	?
R0010 – IPTG Inducible	?

For my research, I decided to focus on acquiring data to characterize and compare the RBSs and promoters. I wanted to create a better equation to allow scientists to predict promoting unit strength when selecting RBSs and promoters separately, as would be typical for a scientist using the Registry. I also created a novel promoter system, borrowed from *B. subtilis*, and wanted to test its strength in comparison with previously existing promoters. This system in *B. subtilis* drives the transition to natural competency and consists of a transcriptional activator molecular and the promoter it identifies. The transcriptional activator is called ComK, which recognizes a sequence called a K-box

($A_4N_5T_4N_xA_4N_5T_4$ where X can either be 8, 18, or 31) (Hamoen 1998). One mRNA transcript of ComK causes 3.6 ± 0.7% of cells to achieve competency (Maamar 2007, Süel 2006), while twenty mRNA transcripts of ComK causes 100% of cells to enter a competent state (Süel 2007). I wanted to see if inserting this promoter and activator from *B. subtilis* into *E. coli* would change its functional capabilities.

Methods

Fluorescence data was taken using a FLx800 microplate fluorescence reader, measuring at an excitation of 485/20 and an emission of 528/20. Optical density was taken using an ELx808 at the 595 endpoint. 200μL of sample were used every reading. All samples were prepared from the same batch of LB media and were grown in 2mL of media in a 37°C incubator for 21.25 hours. Arabinose inductions were done using 0.002% w/v arabinose.

Results

Twenty-one constructs were tested for fluorescence and optical density data, two of which were novel systems borrowed from *B. subtilis*'s competency-inducing cellular mechanisms, called pKbox and pKbox + ComK. The promoter and RBS used, trials completed, Fl/OD (rlu OD^{-1}) data averages, and standard deviations are presented in figure 3-1.

To find where significant differences arose in the data, I performed a series of T-tests and determined p-values for various comparisons in the data. This information is presented in table 3-3.

Table 3-3. Comparisons of different promoter:RBS combinations to determine statistical significance. T-tests were designed to determine if the fluorescence / OD average of the promoter:RBS combination in the right hand column was less than the average of the promoter:RBS combination of the left hand column.

Comparison Test – Promoter:RBS Left Column > Right Column		p-value
None:B0030	None:None	0.109
None:B0031	None:None	0.009
None:B0034	None:None	0.287
None:B0031	None:B0030	0.123
None:B0031	None:B0034	0.048
None:B0030	None:B0034	0.272
R0040:B0030	R0040:None	<0.001
R0040:B0031	R0040:B0030	0.010
R0040:B0034	R0040:B0030	<0.001
R0040:B0034	R0040:B0031	<0.001
I13453:B0030	I13453:None	<0.001
I13453:B0031	I13453:None	<0.001
I13453:B0032	I13453:None	<0.001
I13453:B0034	I13453:None	<0.001
I13453:B0031	I13453:B0030	0.371
I13453:B0031	I13453:B0032	0.302
I13453:B0031	I13453:B0034	0.014
I13453:B0030	I13453:B0032	0.473
I13453:B0030	I13453:B0034	0.046
I13453:B0032	I13453:B0034	0.017
I13453:B003* + Ara	I13453:None + Ara	<0.001
I13453:B003* + Ara	I13453:B003*	<0.001
I13453:B0030 + Ara	I13453:B003(1,2,4) + Ara	<0.001
I13453:B0031 + Ara	I13453:B0032 + Ara	<0.001
I13453:B0031 + Ara	I13453:B0034 + Ara	0.003
I13453:B0032 + Ara	I13453:B0034 + Ara	0.206
pKBox:B0032	pKBox:B0032 + ComK	<0.001
pKBox:None	pKBox:B0032 + ComK	<0.001
None: B0034	pKBox:B0032 + ComK	0.017
pKBox:B0032	None:B0031	<0.001
I13453:B0034	pKBox:B0032	0.046
I13453:B0034 + Ara	I13453:B0031	<0.001
R0040:B0030	I13453:B0030 + Ara	<0.001

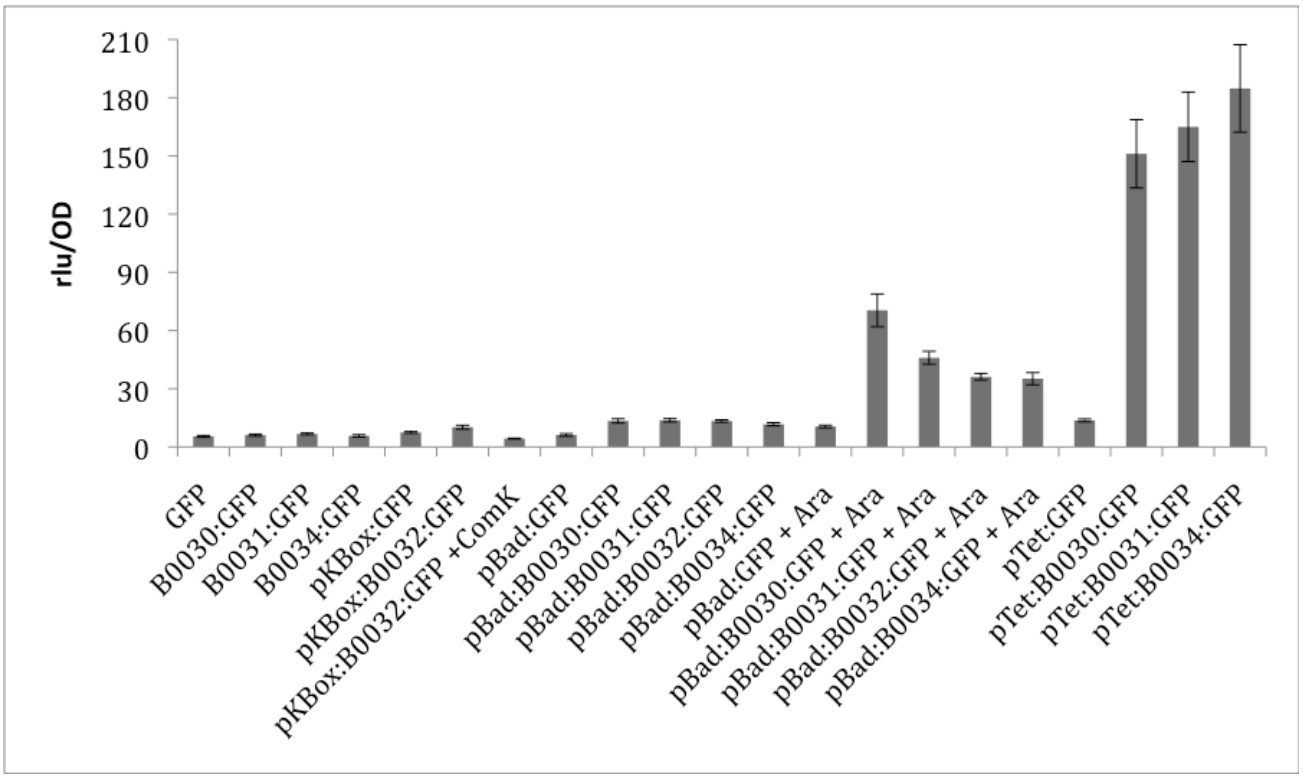

Figure 3-1. The mean and standard deviations of fluorescence / OD data for all 21 constructs. Bars represent one standard error unit.

Discussion

Several points of interest arise when looking at the statistical analysis of the fluorescence data. First of all, except for pKbox with ComK, any promoter causes a significant increase over baseline GFP fluorescence. The only RBS that significantly increases fluorescence above the baseline level without a promoter is B0031.

When considering the promoters as groups, one can see that the RBS that results in the strongest fluorescence in the group of promoters is not consistent. For example, for the arabinose-inducible promoter I13453, RBS B0032 causes the strongest fluorescence, RBSs

B0031 and B0030 are in the middle, and RBS B0034 causes the weakest fluorescence. Alternatively, in the tetracycline-repressible promoter R0040, RBS B0034 causes the strongest fluorescence, then RBS B0031, and RBS B0030 causes the weakest fluorescence. Using various modeling ranking methods, it is possible to ascertain two different rankings for the order of RBS efficiency according to my data (see table 3-4).

Table 3-4. Ranking Preference Sheet, grouped by promoter, ranked by fluorescence / optical density when statistical significance is $p < 0.05$.

Promoter	I13453	I13453 + Ara	R0040	Plain RBS
Highest Rank	B0032	B0030	B0034	B0031
Second Rank	B0031 = B0030	B0031	B0031	B0030
Third Rank	B0034	B0032 = B0034	B0030	B0034

The Borda count method results in B0030 receiving 8 points, B0031 receiving 9 points, B0032 receiving 4 points, and B0034 receiving 6 points, thus ending up with an overall efficiency ranking of B0031 > B0030 > B0034 > B0032. Alternatively, the majority ranking method results in B0031> B0030 > B0032 > B0034. The plurality voting method demonstrates that there is no overall strongest RBS, as a different RBS is the strongest in each promoter comparison (Maki 2006). My data does not demonstrate that one ranking system for strength of RBSs across all promoters can be used. One would expect that if in one group of promoters a certain RBS is the strongest, that trait should transfer when the RBS is paired with other promoters. The gathered data does not indicate that one RBS is always the strongest, no matter which promoter it is with. RBSs in the Registry are based on this premise and are assigned efficiency values ranging from 0.07 (for B0031) to 1 (for B0034), all of which are based on working in conjunction with a promoter activated by the protein *cI* (Weiss 2001). My data dispute these efficiencies, as B0031 is the strongest of the RBSs I tested in two different ranking methods, and B0034 is actually the weakest

according to one ranking method. My data also disputed the idea of constant RBS efficiencies and instead demonstrate that RBSs used in combination with different promoters have varying results. RBS strengths that vary depending on the promoter they are paired with make it very difficult to separate promoter and RBS strength into two separate variables in a modeling equation, so it is easy to see why previous efforts have lumped promoters and RBS into a promoting unit variable – it makes for a simpler model (Leveau 2009), and it allows for varying RBS strength to be overlooked by focusing on the promoter, which causes more dramatic changes.

The promoter data clearly shows that of the promoters tested, R0040 is the strongest, then I13453 + arabinose, then I13453 by itself, then pKBox, then the plain RBSs, and finally the pKBox promoter paired with ComK. Since the choice of promoter has a significant and dramatic influence over the strength of the promoting unit, and the choice of RBS appears to have a small effect as long as an RBS is present, researchers should focus on promoter strengths when making decisions about the level of protein product they desire. More data should be gathered and added to the Registry about promoter and RBS efficiency.

In *B. subtilis*, ComK acts as an activator of genes that follows a K-Box promoter. In *E. coli*, however, the fluorescence of GFP immediately following pKBox is significantly lower than GFP with no promoter or RBS in front of it, indicating that the mechanism of ComK activation is somehow interrupted in *E. coli*. Future research could look at using different K-box promoter sequences as a ComK binding site and see how those different promoters affected fluorescence. *B. subtilis* researchers should look at the possibility of other proteins being involved in activating late competency genes. Other proteins may bind to ComK in *B.*

subtilis that are not present in *E. coli* that cause ComK to act in the cell as a transcriptional activator. It is equally probable that proteins in *E. coli* that are not present in *B. subtilis* downplay the effect of ComK in some way. Previous studies have used physical and in silico approaches to identifying protein interactomes (Ramadan 2008), and may be the most helpful in discerning what proteins bind to ComK. Determining ComK's interactome may be the most helpful first step in figuring out why ComK seemingly switches functions from an activator to a repressor in *E. coli*.

Acknowledgements

Thanks to my academic advisors, Dr. A. Malcolm Campbell and Dr. Laurie J. Heyer. They have both assisted me in countless ways with my research and schoolwork, and also with guidance and support to achieve my goals outside of the classroom.

Thanks to the Center for Interdisciplinary Studies advisor Dr. Scott Denham, and secretaries Linda Shoaf and Vicki Heitman for being patient, understanding, and extremely helpful resources.

Thanks to Will DeLoache, Mike Waters, Pallavi Penumetcha, Kin Lau, Kelly Davis, Erin Feeney, and James Barron for offering advice and constructive criticism at weekly lab meetings. Thanks also to Dr. Drew Endy, Dr. Christina Smolke, Dr. Todd Eckdahl, Dr. Jeff Poet, Dr. Johan Leveau, Dr. Reshma Shetty, and Brian Chow for suggesting ideas at my research presentations or for answering my questions via email.

I would like to thank the funding sources for my research: the Mimms Fellowship from the Davidson Research Initiative, Davidson's James G. Martin Genomics Program, and NSF grant DMS-0733952.

Finally, thanks to my friends and family who talked to me when I was in lab early mornings or late nights, came to my public presentations, and kept me sane throughout this process.

References

Alper, H, C Fischer, E Nevoigt, G Stephanopoulos (2005). Tuning genetic control through promoter engineering. PNAS, 102.36:12678-83.

Andersen, JB, C Sternberg, L Kongsbak-Poulsen, S Petersen-Bjorn, M Givskov, S Molin (1998). New unstable variants of green fluorescent protein for studies of transient gene expression in bacteria. *Applied and Environmental Microbiology* 64: 2240-2246.

Cameron, A.D.S, and R.J. Redfield (2006). Non-canonical CRP sites control competence regulons in *Escherichia coli* and many other γ-proteobacteria. *Nucleic Acids Research,* 34, 6001-14.

Dubnau D, Cirigliano C (1972). Fate of transforming DNA following uptake by competent *Bacillus subtilis*. Journal of Molecular Biology, 64, 9-29.

Dubnau D, Cirigliano C (1974). Uptake and integration of transforming DNA in *Bacillus subtilis*. In Mechanisms in Recombination, ed. RF Grell. New York: Plenum.

Dubnau, D (1991). Genetic Competence in *Bacillus subtilis*. *Microbiological Reviews*, 55, 395-424.

Dubnau D (1999). DNA uptake in bacteria. *Annual Review of Microbiology* 53, 217-244.

Elowitz, MB, and S Leibler (2000). A synthetic oscillatory network of transcriptional regulators. *Nature* 430:335-38.

Grossman, A.D. (1995). Genetic networks controlling the initiation of sporulation and the development of genetic competence in *Bacillus subtilis*. *Annual Review of Genetics*, 29, 477-508.

Gunby P (1978). Bacteria directed to produce insulin in test application of genetic code. *Journal of the American Medical Association* 240.16:1697-78.

Hamoen, L.W., A.F. Van Werkhoven, J.J.E. Bijlsma, D. Dubnau, G. Venema (1998). The competence transcription factor of *Bacillus subtilis* recognizes short A/T-rich sequences arranged in a unique, flexible pattern along the DNA helix. *Genes and Development*, 12, 1539-50.

Hamoen, L.W., B. Haijema, J.J. Bijlsma, G. Venema, and C.M. Lovett (2001). The *Bacillus subtilis* competence transcription factor, ComK, overrides LexA-imposed transcriptional inhibition without physically displacing LexA. *Journal of Biological Chemistry*, 276, 42901-07.

Hamoen, L.W., W.K. Smits, A. de Jong, S. Holsappel, and O.P. Kuipers (2002). Improving the predictive value of the competence transcription factor (ComK) binding site in *Bacillus subtilis* using a genomic approach. *Nucleic Acids Research*, 30, 5517-28.

Hamoen, L.W., G. Venema, and O.P. Kuipers (2003). Controlling competence in *Bacillus subtilis*: shared use of regulators. *Microbiology*, 149, 9-17.

Hanahan, D. Techniques for transformation of *E. coli*. DNA Cloning: A practical approach. Ed. D.M. Glover. IRL Press, 1985, Washington DC.

Johnsborg, O., V. Eldholm, and L.S. Håvarstein (2007). Natural genetic transformation: prevalence, mechanisms and function. Research in Microbiology, 158, 767-778.

Johnson CM, RF Schleif (1995). In vivo induction kinetics of the arabinose promoters in *Escherichia coli*. *Journal of Bacteriology* 177.12:3438-42.

Keasling, Jay (2006). Production of the antimalarial drug precursor artemisinic acid in engineered yeast. *Nature* 440: 940-943.

Knight, Tom (n.d). Idempotent vector design for standard assembly of BioBricks. Accessed 22 Feb. 2009. Available <http://web.mit.edu/synbio/release/docs/biobricks.pdf>.

Leveau, J.H.J, S.E. Lindlow (2001). Predictive and interpretive simulation of green fluorescent protein expression in reporter bacteria. *Journal of Bacteriology*, 183.23:6752-62.

Leveau, Johan (2009). Personal Communication. E-mail 19 February 2009.

Losick, R. and C. Desplan (2008). Stochasticity and cell fate. Science, 320, 65-68.

Lutz, R, and H Bujard (1997). Independent and tight regulation of transcriptional units in *Escherichia coli* via the LacR/O, the TetR/O and the AraC/I1-I2 regulatory elements. *Nucleic Acids Research* 25.6:1203-10.

Maamar, H., R. Arjun, and D. Dubnau (2007). Noise in gene expression determines cell fate in *Bacillus subtilis*. Science, 317, 526-29.

Maki, D and M. Thompson. Selected Case Studies: Social Choice. Mathematical Modeling and Computer Simulation. Thomson, 2006, Belmont, CA.

National Center for Biotechnology Information (2008). GenBank. Accessed 21 Feb. 2009. Available < http://www.ncbi.nlm.nih.gov/Genbank/index.html>.

Palchevskiy, V. and S.E. Finkel (2006). *Escherichia coli* competence gene homologs are essential for competitive fitness and the use of DNA as a nutrient. *Journal of Bacteriology*, 188, 3902-10.

Ramadan, E, et al (2008). Physical and *in silico* approaches identify DNA-PK in a Tax DNA-damage response interactome. *Retrovirology* 5.92.

Reddy, M and J Gowrishankar (2000). Characterization of the uup locus and its role in transposon excisions and tandem repeat deletions in *Escherichia coli. Journal of Bacteriology* 182.7:1978-86.

Registry of Standard Biological Parts (2009). Accessed 21 Feb. 2009. Available < http://partsregistry.org/Main_Page>.

Shetty, RP, D Endy, TF Knight (2008). Engineering BioBrick vectors from BioBrick parts. *Journal of Biological Engineering*, 2.5.

Süel, G.M., J. Garcia-Ojalvo, L.M. Liberman, and M.B. Elowitz (2006). An excitable gene regulatory circuit induces transient cellular differentiation. *Nature* 440, 545-50.

Süel, G.M., RP Kulkarni, J Dworkin, J Garcia-Ojalvo, MB Elowitz (2007). Tunability and noise dependence in differentiation dynamics. *Science*, 315:1716-19.

Van Sinderen, D. and G. Venema (1994). *ComK* acts as an autoregulatory control switch in the signal transduction route to competence in *Bacillus subtilis*. *Journal of Bacteriology*, 176, 5762-70.

Van Sinderen, D. A. ten Berge, B.J. Hayema, L. Hamoen, and G. Venema (1994). Molecular cloning and sequence of *ComK*, a gene required for genetic competency in *Bacillus subtilis*. *Molecular Microbiology*, 11, 695-703.

Weiss, Ron (2001). Cellular computation and communications using engineered genetic regulatory networks. Massachusetts Institute of Technology Ph.D. Thesis. Accessed 27 Mar. 2009. Available < http://www.princeton.edu/~rweiss/papers/rweiss-phd-thesis.pdf>.

Wenhua, L., X. Haiyan, X. Zhixiong, L. Zhexue, O. Jianhong, C. Xiangdong, and S. Ping (2004). Exploring the mechanism of competence development in *Escherichia coli* using quantum dots as fluorescent probes. *Journal of Biochemical and Biophysical Methods*, 58, 59-66.

Appendix. Perl code to find K-boxes in *B. subtilis* and *E. coli*

```
#!/usr/bin/perl

use strict;
use warnings;


#########################################################################
#Enter the length of the sequence you are looking for here
#my $KboxLength = 34; #small
#my $KboxLength = 44; #medium
my $KboxLength = 57; #large

#put the filename of the genome you want to search here - be sure this proGram and the genome file are
        saved in the same folder!
#my $filename = "Ecoli_K12a.fna";
#my $filename = "Ecoli_K12b.fna";
my $filename = "Bsubtilis.fna";

my $genome;
#########################################################################

#This opens the file
open(my $fh, '<', $filename) or die $!;
my @seq = <$fh>;
my $fasta = shift(@seq);
$genome = join("",@seq);
$genome = uc($genome);
$genome =~ s/\s//g;
#print "The genome has ", length($genome), " nt in it \n";

my @sequence;
my @starters;
my @scores;
my $StartSite = 0;
my $z = 0;

#This scores the strand of DNA by adding one point for every base pair that matches the K-box sequence
        exactly, and remembers the sequence if the score is more than 13, making sure that each half of
        the recognized dimer has at least six exact matching base pairs.
print "Match No. \t Score \t Starting bp \t Sequence \n";
while ($StartSite <= (length($genome)-$KboxLength)) {
   my $sequence = substr($genome, $StartSite, $KboxLength);
   my $score = 0;
   my $scoreA = 0;
   my @letter = split(//, $sequence);
   if ($letter[0] eq 'A') { $score = $score + 1;}
   if ($letter[1] eq 'A') { $score = $score + 1;}
   if ($letter[2] eq 'A') { $score = $score + 1;}
   if ($letter[3] eq 'A') { $score = $score + 1;}
   if ($letter[9] eq 'T') { $score = $score + 1;}
   if ($letter[10] eq 'T') { $score = $score + 1;}
   if ($letter[11] eq 'T') { $score = $score + 1;}
   if ($letter[12] eq 'T') { $score = $score + 1;}
```

```
    if ($score >= 6) {
        if ($letter[($KboxLength - 13)] eq 'A') { $scoreA = $scoreA + 1;}
        if ($letter[($KboxLength - 12)] eq 'A') { $scoreA = $scoreA + 1;}
        if ($letter[($KboxLength - 11)] eq 'A') { $scoreA = $scoreA + 1;}
        if ($letter[($KboxLength - 10)] eq 'A') { $scoreA = $scoreA + 1;}
        if ($letter[($KboxLength - 4)] eq 'T') { $scoreA = $scoreA + 1;}
        if ($letter[($KboxLength - 3)] eq 'T') { $scoreA = $scoreA + 1;}
        if ($letter[($KboxLength - 2)] eq 'T') { $scoreA = $scoreA + 1;}
        if ($letter[($KboxLength - 1)] eq 'T') { $scoreA = $scoreA + 1;}
        if ($scoreA >= 6) {
        $score = $score + $scoreA;
        if ($score >= 13) {
            $z++;
            print "($z) $score $StartSite $sequence $KboxLength\n";
        }}}
    else {}
    $StartSite = $StartSite +1;
}


print "Overall, there were $z matches in $filename with more than 13 of 16 matches to a K-box of size
            $KboxLength \n";
```