

## PROMOTER ENGINEERING WITH BIG DATA

(from William and Mary iGEM; contact: [igem@wm.edu](mailto:igem@wm.edu) or [mssaha@wm.edu](mailto:mssaha@wm.edu))

### Summary

In order to give Intro Bio students an understanding of molecular genetics, synthetic biology, and data mining, students will use the iGEM Parts Registry to find related sequences of DNA and assess their function. Promoters are key to transcription and translation, and thanks to synthetic biology, these “parts” have been modularized for easy use in constructing circuits. With the creation of bioinformatics databases and tools like BLAST and GeneBank, it is simple to use known sequences to find related sequences to test as potential novel parts.

### Learning Goals

- a. Students will demonstrate an in-depth understanding of the varying strengths of promoters and their potential applications to real-world problems.
- b. Students will gain an appreciation for the intersection between data science and synthetic biology.
- c. Students will familiarize themselves with sequence analysis, including DNA sequencing technology, sequence databases, and methods to mutate sequences (mutagenesis).
- d. By analyzing sequences, students will review the central dogma of biology.

### Overview

Promoters are regions of a gene located upstream from the coding region that initiate transcription. The sequence of the promoter region determines its characteristics—promoters vary in strength, and some require inducers to activate them. Synthetic biologists can use promoters with different strengths to create different levels of expression in certain genes. These sequences are readily available on databases like NCBI’s GeneBank, where synthetic biology meets big data thanks to genome analysis technology. Furthermore, many synthetic biologists generate vast libraries of sequences by intentionally substituting bases of a sequence (a process known as mutagenesis). In this activity, we will use data science to find related groups of promoters with varying strengths, as well as explore mutations within these promoter

- *What are some applications of strong promoters?*  
In synthetic biology, strong promoters can be used to create a high level of transcription, which would lead to a high production rate of mRNA, and therefore, a high amount of a specific protein.
- *What are some applications of medium strength promoters?*  
In synthetic biology, medium strength promoters can be used to create a moderate level of transcription, which would lead to a moderate production rate of mRNA, and therefore, a moderate amount of a specific protein.
- *What are some applications of weak promoters?*  
Synthetic biologists can use weak promoters to create a low level of transcription, leading to a low level of mRNA production, and therefore a low level of protein production. This is applicable in cases where a high production rate of a protein could harm a cell.

Students will be given three different scenarios that will require various strengths of promoters. After reading each scenario, students will be given examples of promoters with different strengths (based on the scenario) and be asked to find others with a similar strength using the NCBI database. Finally, students will be asked to mutate the promoters they found and use the database to see if their mutated promoters exist in nature.

Scenario 1: Imagine you are a synthetic biologist creating a biosensor. You are engineering a strain of *E. coli* that can decompose the pollutant benzene. In order to decompose a large amount of this pollutant, the bacteria must express a high concentration of the protein XylR, which binds to this pollutant. How strong of a promoter would you use? Use the protocol below to find a promoter with a strength that best matches this situation.

Scenario 2: Imagine you are a synthetic biologist creating an incoherent feedforward loop (IFFL) in *E. coli* based on the mf-Lon protease system. In this IFFL, the mf-Lon protease is an inhibitor that degrades another protein. Although you want the bacterial cells to express mf-Lon, excessive expression contributes to cell toxicity and death. How strong of a promoter would you use? Use the protocol below to find a promoter with a strength that best matches this situation.

Scenario 3: Imagine you are a synthetic biologist transforming a circuit into *E. coli*. In order to visualize the cells, you need to co-transform the bacteria with a second plasmid expressing Green Fluorescent Protein (GFP). However, the expression of genes from the first plasmid places the cells under a high metabolic strain, slowing their growth. You need to transform and express the second plasmid without compounding this strain. Since your microscope can visualize even low levels of fluorescence, very little expression of the fluorescent protein is necessary. How strong of a promoter would you use to express the GFP? Use the protocol below to find a promoter with a strength that best matches this situation.

### Protocol

1. Go to <http://parts.igem.org/Promoters/Catalog/Anderson>.
  - a. Pick a promoter with a strength of:
    - i. 0.60 or greater for strong promoter
    - ii. 0.30-0.60 for medium-strength promoter
    - iii. 0.10-0.30 for weak promoter
  - b. Copy the sequence of the chosen promoter
2. Go to <https://blast.ncbi.nlm.nih.gov/Blast.cgi> and click "Nucleotide BLAST."
3. Paste promoter sequence into box labeled "Enter accession number(s), gi(s), or FASTA sequence(s)."
4. Under "Database," select "Nucleotide collection (nr/nt)."
5. Under "Optimize for," select "Somewhat similar sequences (blastn)," then click "BLAST."
6. Scroll down to "Alignments."
7. Select any sequence that has a percent identity greater than 75% and less than 100%.
8. Copy sequence located next to "Sbjct." This should be a promoter with similar strength to the original one found on iGEM Registry.
9. To calculate the strength of the promoter, go to <http://www.bacpp.bioinfocps.com/home> and enter the sequence copied in the previous step.
10. Then, return to <https://blast.ncbi.nlm.nih.gov/Blast.cgi> and do a "Nucleotide BLAST" for each of the sequences. Make sure to select "Highly similar sequences (megablast)."

11. Record any sequences that have a percent identity of 100%.
12. Answer the scenario questions.