Algorithms for designing genetically stable multi-copy circuits

Authors: Moritz Burghardt, Tamir Tuller

Synthetic biology aims to maximize the expression levels of target proteins. A common strategy is integrating multiple copies of a target gene into a host organism's genome. However, repetitive sequences can induce homologous recombination, reducing the copy number of a target gene. In addition, overproduction of a target protein can cause a metabolic burden, which creates a selection pressure for decreasing copy number. Various studies have reported a strong relationship between shared substring length and homologous recombination rate (Shen 1986, Aw 2013). Therefore, to reduce this risk, nucleotide sequences encoding a target protein should be different and, in particular, not share long common subsequences.

To address this problem, we present an algorithm to design a set of highly expressed coding sequences less likely to induce homologous recombination. The approach is based on the Chimera algorithms, where sequences are composed of blocks that appear in the host genome and therefore adhere to the gene expression rules of the host (Zur 2015). In addition, we propose a metric to estimate the recombination rate between two sequences.

We generate and evaluate gene sets for various hosts and proteins of interest and find that our algorithm designs sequences that are unlikely to induce recombination and well adapted to the expression machinery of the host. The algorithm has been implemented in the ChimeraUGEM software tool available for download on the Tuller lab website.