Introduction

Molecular crowding is a biophysical constraint on the ability of microbial cell factories to produce proteins of interest at sufficient titer, rate and yield to become economically viable. To address this issue directly, the Cellular Capacity Optimization project aims to iteratively knock out nonessential regions of the genome that code for highly expressed proteins in order to reclaim that protein capacity for engineered proteins. In order to rationally guide this process, we have developed `MaxMass`, an algorithm for iterative reduction of the host proteome without replacement. `MaxMass` automatically calculates which genes are essential for the current environment and iteration, and uses these essential genes as constraints on the size of the knockout region, while optimizing for maximum protein mass reclamation.

Methods

Max Mass Objective

The `MaxMass` algorithm contains the following objective function:

$$\max \sum_{k\in K} y_k \sum_{i=1}^{k} m_i - \sum_{k\in K} x_k \sum_{i=1}^{k} m_i$$

where the variables are

$$x_k = \begin{matrix} 1, \text{ if gene or promoter k is the first gene or promoter within the deleted segment} \\ 0, \text{ otherwise} \end{matrix}$$

and

$$y_k = \begin{matrix} 1, \text{ if gene or promoter k is immediately after the end of the deleted segment} \\ 0, \text{ otherwise} \end{matrix}$$

where

$$\sum_{k\in K} x_k = 1$$

and

$$\sum_{k\in K} y_k = 1$$

and the parameters are

- d_k: Position of the first nucleotide of the deleted sequence starting from the origin of replication when gene or promoter k is selected to be deleted in the begining of the stretch. Note that d_k is not always the start site of a gene or promoter. It is the first nucleotide of the nonoverlapped region between the gene/promoter k and gene/promoter $k-1$
- d'_k: Position of the first nucleotide of the gene or promoter k immediately after the deleted sequence
- m_i is the measured protein mass of gene i and $\sum_{i=1}^{k} m_i$ is the cumulative protein mass of all genes from the origin of replication to the kth gene. By subtracting the cumulative protein mass of the start gene (x_k) from the cumulative protein mass of the end gene (y_k), we obtain the cumulative protein mass of the interval for the optimal solution

We maximize the mass that is knocked out based on absolute quantitative proteomics Schmidt et al 2015 (http://www.nature.com/articles/nbt.3418) in (fg/cell)

MaxMass constraints

$\sum_{j\in J} S_{i,j} v_j = 0$	$i = 1, \dots, N$	Steady state conservation of metabolite i requirement
$\sum_{k\in K} y_k = 1$	$\sum_{k\in K} x_k = 1$	Ensure only one gene knockout per iteration
$\sum_{j=1}^{k} x_j - \sum_{j=1}^{k} y_j = z_k$	$k = 1, \dots, K$	Ensures all genes between start and end area also knocked out
$v_{\text{biomass}} \geq f \cdot v_{\text{biomass, max}}$		Requires that the KO must be capable of growing at the specif
$z_g = \prod_{p\in p^g} z_p$	$p = 1, \dots, P$	Ensures that a gene is not functional if all of its promoters are

Gene Protein Reaction constraints

Single gene catalyzes reaction: $\mathrm{k} \rightarrow \mathrm{v_j}$

$$(1 - z_k) \cdot LB \leq v_j \leq (1 - z_k) \cdot UB$$

Enzyme dimer complex catalyzes reaction: $(\mathrm{k_1\ AND\ k_2}) \rightarrow \mathrm{v_j}$

$$(1 - z_{k_1}) \cdot LB \leq v_j \leq (1 - z_{k_1}) \cdot UB$$
$$(1 - z_{k_2}) \cdot LB \leq v_j \leq (1 - z_{k_2}) \cdot UB$$

Two Isozymes catalyze reaction: $(\mathrm{k_1\ OR\ k_2}) \rightarrow \mathrm{v_j}$

$$(2 - z_{k_1} - z_{k_2}) \cdot LB \leq v_j \leq (2 - z_{k_1} - z_{k_2}) \cdot UB$$
$$LB \leq v_j \leq UB$$

Complex Gene Protein Reaction: $(\mathrm{k_1\ AND\ k_2})\ \mathrm{OR}\ (\mathrm{k_1\ AND\ k_3}) \rightarrow \mathrm{v_j}$

$$(2 - z_{k_2} - z_{k_3}) \cdot LB \leq v_j \leq (2 - z_{k_2} - z_{k_3}) \cdot UB$$
$$(1 - z_{k_1}) \cdot LB \leq v_j \leq (1 - z_{k_1}) \cdot UB$$

Results

Comparison of Predictions with Experiment

To evaluate the ability of the algorithm to predict protein reclamation, we used proteomics data obtained from the KHK collection of knockouts. As a baseline first step, we wished to assess how well the protein distribution of the previous step was able to predict the protein distribution of the subsequent step by assuming that the only expression that changed between steps were the genes that were knocked out. The results of this analysis are displayed in Figure XX. Interestingly, there were some proteins whose expression level increased as a result of the knockout.

Application of `MaxMass` to wild-type E. coli W3110

We then applied `MaxMass` using the protein expression from wild-type E. coli strain W3110 Schmidt et al 2018 (https://www.nature.com/articles/nbt.3418) in order to predict which gene knockouts would result in the largest amount of protein reclamation. The results for this experiment are displayed in Figure YY. We predicted that ZZ% of the proteome could be reclaimed after N steps.

Discussion and Future work

One interesting observation is that knocking out the most highly expressed proteins can result in new highly expressed proteins. To predict which proteins **must** be overexpressed in the knockout, future work will focus on extending the metabolic model with RNA and Protein expression mechanisms, known as ME-models. We will also extend this approach for other organisms such as **Pseudomonas flourescens**, **Shewanella oneidensis** MR-1 and **E. coli** Nissle.