Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Biotechnology

Improving the success and impact of the metabolic engineering design, build, test, learn cycle by addressing proteins of unknown function

Laura R Jarboe[1,2]

Rational, predictive metabolic engineering of organisms requires an ability to associate biological activity to the corresponding gene(s). Despite extensive advances in the 20 years since the *Escherichia coli* genome was published, there are still gaps in our knowledge of protein function. The substantial amount of data that has been published, such as: omics-level characterization in a myriad of conditions; genome-scale libraries; and evolution and genome sequencing, provide means of identifying and prioritizing proteins for characterization. This review describes the scale of this knowledge gap, demonstrates the benefit of addressing the knowledge gap, and demonstrates the availability of interesting candidates for characterization.

Addresses
[1] Chemical and Biological Engineering, Iowa State University, Ames, IA 50011, United States
[2] Interdepartmental Microbiology Graduate Program, Iowa State University, Ames, IA 50011, United States

Corresponding author: Jarboe, Laura R (ljarboe@iastate.edu)

Current Opinion in Biotechnology 2018, **53**:93–98

This review comes from a themed issue on **Chemical biotechnology**

Edited by **Patrick Cirino** and **Mattheos Koffas**

https://doi.org/10.1016/j.copbio.2017.12.017

Introduction

Metabolic engineering is, as with other types of engineering, based on a rational, predictive approach. In recent years, the metabolic engineering approach has been frequently described in the 'design, build, test, learn' framework (Figure 1) [1•,2]. The design phase, particularly when the intention is to eliminate certain reactions, requires an ability to attribute biochemical activity to the associated genes. However, even the most commonly utilized model microbial cell factories still have proteins with unknown function. Put more bluntly — how can you effectively manage and improve a production facility when you do not know what all of the employees are doing? This incomplete knowledge of the source of functionality also limits the design space when using a biofoundry approach [3,4] and cell-free biosynthesis [5,6], as enzymes of unknown function cannot be included in the design of these systems. This review gives an overview of the problem of incomplete protein annotation and the benefits of addressing this knowledge gap.

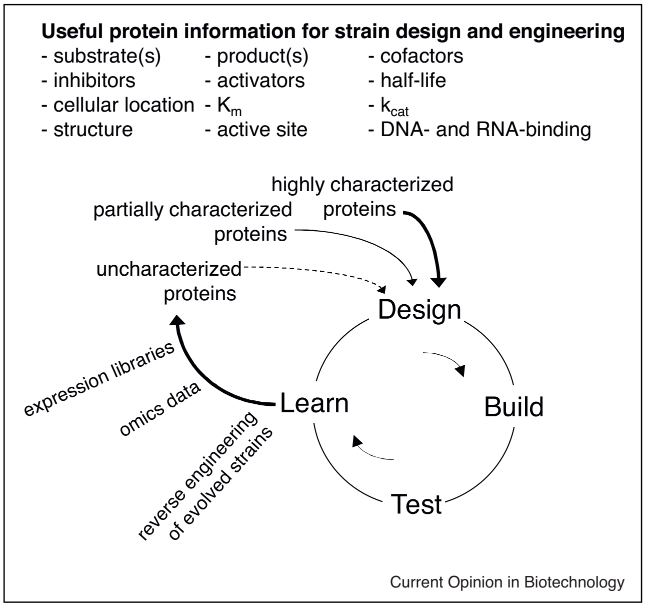
The prevalence and problem of unannotated proteins

The first *E. coli* genome consisted of 4288 protein-coding genes of which 1909 (44.5%) had functional classification 'putative enzyme', 'other known genes' or 'hypothetical, unclassified, unknown' [7]. Ten years later, 1077 genes either entirely lack or have only a partial assigned biochemical function [8]. This decrease in the number of genes lacking an assigned function demonstrates both a substantial improvement in our understanding of this model organism and our ability to predict enzyme function. However, the knowledge gap is still non-trivial: a 2014 analysis concluded that among the 2048 domains associated with 4210 *E. coli* proteins, 359 were of unknown function [9]. Currently, EcoCyc provides publicly available Smart Tables featuring genes with minimal characterization [10].

A clear demonstration of how this knowledge gap impacts our ability to rationally design organisms is Hutchison *et al.*'s successful minimization of the *Mycoplasma mycoides* genome from the already small size of 1080 kbp and 901 genes to 531 kbp and 473 genes [11••]. Their first pass in this genome minimization was deletion of 440 supposedly non-essential genes based largely on existing biochemical knowledge. This rational design did not result in a viable microbe. The subsequent successful genome minimization strategy consisted of a recursive process of transposon mutagenesis and classification of genes as essential, non-essential and quasi-essential. Surprisingly, approximately 30% of the genes in this final minimal genome have unknown biological function. Put more bluntly, our current efforts to generate the smallest functional genome required the inclusion of a very large number of genes that we do not understand their purpose.

Another demonstration of how our knowledge gap impacts organism design is Forsberg *et al.*'s use of a metagenomics library from soil to identify genes conferring tolerance to biomass-derived inhibitors [12]. Two of the most promising genes, one of which increased tolerance to ferulic acid and the other to furfuryl alcohol, have

Figure 1



Our ability to engineer microbial cell factories for improved performance via the design, build, test, learn (DBTL) cycle depends on the availability of knowledge regarding the associated proteins. Uncharacterized proteins are an underutilized resource in metabolic engineering applications. Increasing focus on the learning component of the metabolic engineering cycle, particularly in the context of these uncharacterized genes and proteins, can add value and impact to existing and forthcoming datasets and decreases the magnitude of the knowledge gap.

unknown function. In order to learn from this successful iteration of the 'design, build and test' cycle, we need to have some insight into what these proteins are doing. This knowledge would then provide additional insight into the mechanism of toxicity for these compounds and inspire design strategies for further strain improvement.

Beyond impacting the fields of metabolic engineering and synthetic biology, this knowledge gap is also relevant to public health. A recent analysis of carbapenem-resistant *Enterobacteriaceae* identified plasmids that enabled carbapenem resistance, though this resistance could not be explained by current understanding of any of the plasmid-encoded genes [13•]. An inability to attribute the carbapenem resistance to a specific gene interferes with the ability to identify this resistance phenotype from sequence data.

This knowledge gap not only impacts our understanding of microbial metabolism, physiology and pathogenesis, but is also a problem with higher-order model organisms, such as Arabidopsis [14].

Characterization of gene essentiality

Hutchison's genome minimization project ultimately relied on experimental assessment of gene essentiality to guide the selection of genes for deletion [11••]. Of the 359 domains of unknown function in the *E. coli* proteome, 89 are classified as essential [9]. Identification of conditions in which a focal gene is essential, or in which its absence significantly impairs growth, not only guides metabolic engineering strategies [15], but also provides some insight into its function [16].

The Keio collection targeted 4288 genes in *E. coli* BW25113 for in-frame deletion [17,18]. Eighteen of the 315 genes found to be essential in the Keio collection were in COG categories S, U or V — function unknown. Now, more than 10 years after this initial report, only two of these genes lack a fully characterized function (Table 1). Of the 16 essential genes that now have a characterized function, several have been used in metabolic engineering strategies to improve strain performance (Table 1). This demonstrates the benefit associated with enzyme characterization.

Knowledge about gene essentiality also plays a role in assessing the performance of genome-scale models, though this approach is only relevant to genes with sufficient biochemical characterization for inclusion in a

Table 1

Keio collection essential genes that were in COG categories S, U or V ('function unknown') when [17] was published. Gene synonyms and activity descriptions are from EcoCyc 21.1, September 2017. Not shown: *erpA*, *ftsY*, *lepB*, *lptA*, *lptC*, *lpxH*, *lspA*, *secA*, *secD*, *secE*, *secF*, *secY* and *yefM*

Gene name	Activity	Representative application for strain improvement
Select genes that now have assigned biochemical function and have been used for improvement of strain performance		
msbA	ATP-dependent lipid A-core flippase	Increased expression led to a greater than twofold increase in amorphadiene production [46]
ffh	Signal recognition particle protein component	Increased expression led to >3-fold increase in abundance of immunoglobulin G (IgG) [47]
yidC	Membrane protein insertase	Increased yield of recombinant membrane protein sensory rhodopsin II [48]
Genes still lacking assigned biochemical function		
yhhQ	Putative queuosine precursor transporter	
yigP, now *ubiJ*	Ubiquinone biosynthesis accessory factor	