Capturing Full Cellular Regulation In silico using “Big” Data: A Frontier for Systems Biology Perspectives

Steven Sowa¹, Jorge Vazquez-Anderson² and Lydia M Contreras¹,²*

¹Microbiology Graduate Program, University of Texas at Austin, 100 E. 24th Street, Austin, Texas 78712, USA
²McKetta Department of Chemical Engineering, University of Texas at Austin, 200 E. Dean, Keeton St., Stop C0400, Austin TX 78712, USA

This perspective article offers our view on the current and future directions of the integration of “big” data and genome-wide engineering. In our perhaps not-so-distant view of the future, a desired phenotype can be simply envisioned and inputted into a computational algorithm to obtain a detailed experimental strategy that would make it happen. It is foreseeable that a detailed map of a fully integrated regulatory and metabolic network can be generated using already built-in capabilities, resulting from large-scale genomics, proteomics, transcriptomics and metabolomics data.

Two important questions that arise from genome-wide engineering approaches are: how can we systematically search the genome for targets that do something we care about? And how do we achieve predictable system-wide tunability of gene expression? As part of these answers, systems biology approaches have most recently turned to cellular regulators.

Already, the combination of systems-wide experimental approaches and mathematical modeling has allowed new ways of thinking about controlling and optimizing large-scale gene expression. In particular, the use of modeling tools supports the central vision of large scale genome engineering: that optimal gene targeting schemes can be determined a priori to allow rational synthesis of specific patterns that can best contribute to a desired trait. Recently, systems approaches to genome regulation have echoed “big data” approaches in biology, where a tremendous focus has been placed on the simultaneous large-scale characterization of all cellular effects (e.g. proteomics, transcriptomics, metabolomics) [1-4]. A vision of where the complete merging of computational and experimental systems approaches could lead us is depicted in Figure 1.

The emphasis on the use of large biological data sets in strain engineering is evidenced by a number of experimental genome-wide engineering strategies that have recently emerged to rapidly evolve specific metabolic functions in the context of all natural metabolic pathways. These approaches (e.g. MAGE, CAGE, and TRMR [5-7]), are briefly described below and target both the coding content of the genome, as well as multiple promoter regions to introduce genome-wide modifications that improve cellular fitness.

Multiplex Automated Genome Engineering (MAGE) is capable of attaining genomic diversity by simultaneously introducing mutations in many locations of the genome in a single cell or across populations. In this way, MAGE rises as a cutting-edge technique by accelerating the evolution of improved metabolically relevant strains. This method allows for the automated large-scale programming, and evolution of cells and has been showcased by applying oligo-mediated allelic replacement in E. coli. One end goal of this approach is the optimization of metabolic pathways, with the ultimate purpose of overproducing industrially relevant components [5]. More Recently, hierarchical Conjugal Assembly Genome Engineering (CAGE) has been applied in the development of genome-wide replacement of all TAG for TAA stop codons in parallel across 32 E. coli strains [6]. Hierarchical Conjugal Assembly Genome Engineering (CAGE), MAGE’s more powerful sibling, enables the recombination of genomic modifications in pairs by hierarchically transferring the codon deletions from a donor cell to a recipient cell in a series of successive conjugations. Remarkably, this method can be applied so that all 314 stop codon modifications can be introduced into a single fully recoded strain. Ultimately, CAGE arises as a complementary method to the proven ability of MAGE to introduce nucleotide-scale modifications across the genome and allows for the in vivo assembly of modified chromosomes. While MAGE and CAGE enable the large-scale genomic modifications of relevant genes, these approaches assume that the targeted locations are known a priori (Box 1).

Tractable Multiplex Recombineering (TRMR) [7] is another genome-wide methodology that combines multiplex DNA synthesis [8-12], recombineering [13-15] and barcoding technology [16,17], for the simultaneous mapping of genetic mutations and their corresponding traits. Application of this method has allowed perturbation of the expression levels of >95% of genes in E. coli by introducing DNA cassettes and barcode sequences upstream each gene. In general, a major breakthrough has been the ability to map thousands of genes in several conditions via the use of barcoded and microarray technologies. It is also worth noting that a series of other significant efforts have preceded the genome engineering methodologies summarized above. Others have included whole genome assembly [18], developing a “minimization” method in which large segments of unstable DNA are eliminated in the genome [19], and transforming entire genomes across microorganisms [20]. The Church group has written a relatively recent review of these genome engineering techniques [21]. This wave of large-scale combinatorial and evolutionary methods to engineer entire biological systems has been enabled by major advancements in DNA synthesis tools and by techniques for manipulating, synthesizing and recombining DNA, in an almost a la carte manner [21].

These experimental approaches have brought us closer to the dream of simultaneously targeting entire genomes for fast evolution. This has been in part due to the realization that creating complex phenotypes requires simultaneous manipulations of multiple genes [22,23]. These systems approaches have been highly justifiable by the understanding that control and regulation of cellular metabolism is distributed over multiple enzymes, and that multiple mutations are
### Method

<table>
<thead>
<tr>
<th>Method</th>
<th>Designers</th>
<th>Year</th>
<th>Purpose</th>
<th>Basic Steps</th>
<th>Metabolic Model</th>
<th>Constraints</th>
<th>Objectives</th>
<th>Optimization Variables</th>
<th>Intermediate Result</th>
<th>Results</th>
<th>Strengths</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conjugative Assembly Genome Engineering (CAGE)</td>
<td>Wang et al.</td>
<td>2011</td>
<td>An experimental technique to generate sequence diversity on a genomic level in an automated and targeted fashion.</td>
<td>A. Multiplex Assembly Genome Engineering (MAGE) 1. Grow cells, express recombineering proteins. 2. Synthesize a custom degenerate oligo library that targets a variety of chromosomal loci. 3. Electroporate oligo library into <em>Escherichia coli</em>. 4. Recover cells and repeat cycling.</td>
<td>Flux balance metabolic model (iMC101025) with boolean logic to control transcription regulation and gene expression.</td>
<td>Limited number of gene deletions and overexpressions</td>
<td>Maximizes metabolite production and growth</td>
<td>Levels of enzymes (delete or overexpress enzymes)</td>
<td>After cycling through MAGE, greater genetic diversity in ~10-25 loci has been created.</td>
<td>Strain with greater diversity at hundreds of loci if needed.</td>
<td>Targeted and automated method for creating genetic diversity.</td>
</tr>
</tbody>
</table>

**Box 1:** Experimental and computational methods representative of the tools available for genome engineering.

---

**Figure 1:** The “dream” of *In silico*-aided genome-wide engineering.

- Real-time feedback
- Targets & genome engineering strategy
- Evaluation of biologically achievable results
- Living system
- Desirable phenotype

**In silico**

Mapping of integrated regulatory and metabolic network

1. Identification of targets
2. “Game plan” for phenotype construction
3. Simulation of outcome

**Big data:** genomics, proteomics, transcriptomics, modomics.

**Genome-wide engineering**
required to alter expression even of a single enzyme [24]. Importantly, what these approaches have in common is the emphasis on rationally creating the shortest evolutionary path to a desired trait. These techniques, in essence, try to rewrite the many synergistic, regulatory and feedback effects present in cellular circuity. These methods also aim to modify levels of enzymes involved in multiple enzymatic pathways (e.g. multiple knockouts that can redirect metabolic flux) for greater perturbation of metabolic behavior. However, to date, most genome-wide approaches have been demonstrated in the context of model organisms. A challenge moving forward will be to showcase large-scale strategies to control global gene expression in a targeted way in other organisms besides *E. coli* and *Saccharomyces cerevisiae*.

In addition, these methods operate *ad hoc*, targeting a vast number of enzymes and proteins that might not be functionally related to a desired phenotype. A major risk with these approaches is the tradeoff between achieving more diversity and perturbing larger gene sets as the latter can interfere with function; this is especially the case, as these strategies can represent uncoordinated genome modifications. In this case, a major challenge is the risk of deteriorated strain performance, given that resulting metabolic configurations do not take into account optimal interdependence of the affected pathways.

To help predict beneficial genome targets that could be tuned simultaneously to produce optimal phenotypes, several genome-wide metabolic models and optimization frameworks have been constructed [25]. Importantly, a series of optimization methods have been developed [26-30], to formulate strategies *a priori* for rerouting metabolites by controlling gene expression in a highly rational way. As an example, the Maranas’ lab recently developed a cellular optimization framework called OptForce (an improvement to previous work with OptKnock and OptReg). This approach uses flux data from wild type cells to determine which genes need to be up–or down-regulated by identifying fluxes that would have to change significantly relative to wild type, in order to achieve a metabolic objective [26]. Although similar to OptKnock, OptORF (Box 1) has been developed to specifically account for potential manipulation of transcriptional regulation [29]. In an attempt to further contribute to the modeling of regulation, a different group has developed a flux scanning technique, based on enforced objective flux (FSEOF) to maximize a biomass objective [31]. Thus far, these simulations have been used to identify reactions (and, therefore, gene targets) that have large shifts in flux when product formation is high. It is encouraging that these (and other similar) methods have aided the rational design of several metabolite producing strains [32,33].

As we ponder upon where to go next with guiding genome-wide regulation, the “Utopia” of genome-wide engineering becomes an important frame of reference. That is, we would love for computational systems models to lay the path towards tapping into relevant patterns of gene expression that are actually important to the function in question. Recent collaborative databases such as K-base and subtiwiki [34], highlight current interest in dovetailing experimental and computational approaches into powerful engineering tools. The ability to obtain a coherent genome-wide engineering game plan “from the get-go” will likely offer an important advantage over the large-scale regulation of unsynchronized regulators (e.g. unrelated transcription factors), by random library approaches. In addition, experimentalists continue to envision several abilities that include: targeting the minimal number of molecules to induce significant strain diversity, simultaneously managing functionally diverse pathways and preventing disruption of any other cellular activities by isolating the engineering of an individual metabolic function.

So what can computational systems approaches do? The prediction of organism-wide impact of regulators and variants thereof on a specific phenotype requires thorough quantitative understanding of the expression of the regulating entities, in the context of specific intra- and extra-cellular conditions. Moreover, one requires a clear map of the effect of changing these regulators on intracellular metabolic fluxes, proteins and mRNA transcript levels. However, mathematical understanding of cellular regulation is in its incipient stages [35,36]. Given that current (metabolic flux and kinetic) models do not explicitly reflect the mechanistic influence of any form of gene regulation, full predictive capabilities for deciphering which regulators to target do not yet exist. Yet, it would be remarkable to determine genomic targets *a priori* to create desired diversity for strain customization based on a desired optimization objective. Such a vision of *in silico*-aided genome engineering is depicted by Figure 1. In the perhaps not-so-distant view of the future, a desired phenotype can be simply envisioned and inputted into a computational algorithm to obtain a detailed experimental strategy that would make it happen. It is foreseeable that a detailed map of a fully integrated regulatory and metabolic network can be generated using already built-in capabilities, resulting from large-scale genomics, proteomics, transcriptomics and modomics data. In this way, it would be highly feasible to obtain (1) potential molecular and/or pathway targets, (2) genome engineering strategies and (3) simulation of how genome modifications would play out in a biologically-relevant way. Importantly, these would offer a highly guided strategy for executing effective systems-wide engineering at the bench. Moreover, this includes the vision of an iterative process where experimental data obtained from phenotypic evaluations will be used for continual algorithm improvement. If we further stretch our imagination of the future, it is highly possible that, before too long, such an integrated *in silico*-experimental setup can operate in real time. It is even more exciting to consider the possibility of a feedback closed-loop system that would continually optimize a target living system for a desired phenotype base on generated data. Rapid progress is already being made towards complete integration of large-scale experimental and computational efforts that target both metabolic and regulatory pathways, although we are only at an early stage.

**Acknowledgement**

We are grateful for funding to L.M.C. from the Welch Foundation (Grant NO. F-1758), Defense Threat Reduction Agency (DTRA) Young Investigator Program (HDRTA1-12-0016), Air Force Office of Scientific Research (AFOSR) Young Investigator program (FA9550-13-1-0160), and the NSF CAREER program (CBET-1254754). We want to acknowledge as well the National Science and Technology Council of Mexico for the graduate fellowship granted to J.V.A (CONACYT-194638). We also thank Kevin Baldridge for his valuable comments and edition contributions towards the successful completion of this manuscript.

**References**


