Current Synthetic and Systems Biology

Sowa et al., Curr Synthetic Sys Biol 2013, 1:1
DOI: 10.4172/2332-0737.1000107

Short Communication Open Access

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

Steven Sowa[1], Jorge Vazquez-Anderson[2] and Lydia M Contreras[1,2]*

[1]*Microbiology Graduate Program, University of Texas at Austin, 100 E. 24th Street, Austin, Texas 78712, USA*
[2]*McKetta 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 modomics 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, modomics) [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 compounds [5]. More Recently, hierarchical Conjugative 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 Conjugative 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

***Corresponding author:** Lydia M Contreras, McKetta Department of Chemical Engineering, University of Texas at Austin, 200 E. Dean. Keeton St., Stop C0400, Austin TX 78712, USA, E-mail: lcontrer@che.utexas.edu

Received September 10, 2013; **Accepted** October 16, 2013; **Published** October 19, 2013

Citation: Sowa S, Vazquez-Anderson J, Contreras LM (2013) Capturing Full Cellular Regulation *In silico* using "Big" Data: A Frontier for Systems Biology Perspectives. Curr Synthetic Sys Biol 1: 107. doi: 10.4172/2332-0737.1000107

Method	Conjugative Assembly Genome Engineering (CAGE)[6]
Designers	Wang et al.
Year	2011
Purpose	An experimental technique to generate sequence diversity on a genomic level in an automated and targeted fashion.
Basic Steps	A. Multiplex Assembly Genome Engineering (MAGE) 1. Grow cells, express recombineering proteins. 2. Synthesize a custom degenerate oligo library that targets a variety chromosomal loci. 3. Electroporate oligo library into *Escherichia coli*. 4. Recover cells and repeat cycling.
Intermediate Result	After cycling through MAGE, greater genetic diversity in ~10-25 loci has been created.
	B. Conjugative assembly is used to join multiple MAGE evolved strains into a single strain. 1. Pair two independently evolved MAGE strains. 2. F plasmid is used to transfer the MAGE-evolved loci of one strain into the other. NOTE: Fidelity and efficiency of the transfer process is ensured by multiple selectable markers in the donor and recipient.
Results	Strain with greater diversity at hundreds of loci if needed.
Strengths	Targeted and automated method for creating genetic diversity.

Method	OptORF[29]
Designers	Kim and Reed
Year	2010
Purpose	A computational technique to predict the effects of genetic overexpressions and knockouts on metabolism.
Metabolic Model	Flux balance metabolic model (iMC1010[25]) with boolean logic to control transcription regulation and gene expression
Problem	Bilevel constraint-based optimization problem
Objectives	Maximize metabolite production and growth
Optimization Variables	Levels of enzymes (delete or overexpress enzymes)
Constraints	Gene-protein-reaction associations Transcriptional regulation Limited number of gene deletions and overexpressions
Results	Develops a combination of gene deletions and overexpressions that maximize production of a metabolite of interest.
Strengths	Accounts for transcriptional regulation in predictions. Accounts for the all isozymes that control flux through a given reaction.

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

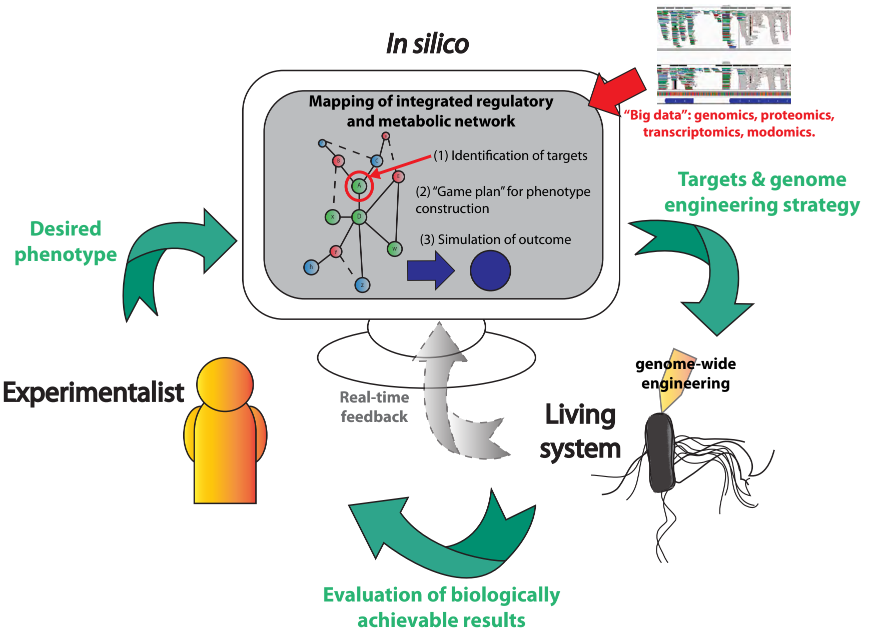
In silico
Mapping of integrated regulatory and metabolic network
(1) Identification of targets
(2) "Game plan" for phenotype construction
(3) Simulation of outcome
"Big data": genomics, proteomics, transcriptomics, modomics.
Targets & genome engineering strategy
Desired phenotype
Experimentalist
Real-time feedback
Living system
genome-wide engineering
Evaluation of biologically achievable results

Figure 1: The "*dream*" of *In silico*-aided genome-wide engineeering.

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 rewire the many synergistic, regulatory and feedback effects present in cellular circuitry. 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-1756), Defense Threat Reduction Agency (DTRA) Young Investigator Program (HDTRA1-12-0016), 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

1. Perrenoud A, Sauer U (2005) Impact of global transcriptional regulation by ArcA, ArcB, Cra, Crp, Cya, Fnr, and Mlc on glucose catabolism in *Escherichia coli*. J Bacteriol 187: 3171-3179.
2. Wessely F, Bartl M, Guthke R, Li P, Schuster S, et al. (2011) Optimal regulatory strategies for metabolic pathways in *Escherichia coli* depending on protein costs. Mol Syst Biol 7: 515.
3. Liang JC, Bloom RJ, Smolke CD (2011) Engineering biological systems with synthetic RNA molecules. Mol Cell 43: 915-926.
4. Na D, Yoo SM, Chung H, Park H, Park JH, et al. (2013) Metabolic engineering of *Escherichia coli* using synthetic small regulatory RNAs. Nat Biotechnol 31: 170-174.
5. Wang HH, Isaacs FJ, Carr PA, Sun ZZ, Xu G, et al. (2009) Programming cells by multiplex genome engineering and accelerated evolution. Nature 460: 894-898.

6. Isaacs FJ, Carr PA, Wang HH, Lajoie MJ, Sterling B, et al. (2011) Precise manipulation of chromosomes *in vivo* enables genome-wide codon replacement. Science 333: 348-353.

7. Warner JR, Reeder PJ, Karimpour-Fard A, Woodruff LB, Gill RT (2010) Rapid profiling of a microbial genome using mixtures of barcoded oligonucleotides. Nat Biotechnol 28: 856-862.

8. Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, et al. (1991) Light-directed, spatially addressable parallel chemical synthesis. Science 251: 767-773.

9. Blanchard AP, Kaiser RJ, Hood LE (1996) High-density oligonucleotide arrays. Biosens Bioelectron 11: 687-690.

10. Singh-Gasson S, Green RD, Yue Y, Nelson C, Blattner F, et al. (1999) Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. Nat Biotechnol 17: 974-978.

11. Cleary MA, Kilian K, Wang Y, Bradshaw J, Cavet G, et al. (2004) Production of complex nucleic acid libraries using highly parallel *in situ* oligonucleotide synthesis. Nat Methods 1: 241-248.

12. Ghindilis AL, Smith, MW, Schwarzkopf KR, Roth KM, Peyvan K, et al. (2007) CombiMatrix oligonucleotide arrays: genotyping and gene expression assays employing electrochemical detection. Biosens Bioelectron 22: 1853-1860.

13. Yu D, Ellis HM, Lee EC, Jenkins NA, Copeland NG, et al. (2000) An efficient recombination system for chromosome engineering in *Escherichia coli*. Proc Natl Acad Sci U S A 97: 5978-5983.

14. Zhang Y, Buchholz F, Muyrers JP, Stewart AF (1998) A new logic for DNA engineering using recombination in *Escherichia coli*. Nat Genet 20: 123-128.

15. Murphy KC (1998) Use of bacteriophage lambda recombination functions to promote gene replacement in *Escherichia coli*. J Bacteriol 180: 2063-2071.

16. Shoemaker DD, Lashkari DA, Morris D, Mittmann M, Davis RW (1996) Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. Nat Genet 14: 450-456.

17. Giaever G, Chu AM, Ni L, Connelly C, Riles L, et al. (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. Nature 418: 387-391.

18. Gibson DG, Benders GA, Axelrod KC, Zaveri J, Algire MA, et al. (2008) One-step assembly in yeast of 25 overlapping DNA fragments to form a complete synthetic *Mycoplasma genitalium* genome. Proc Natl Acad Sci U S A 105: 20404-20409.

19. Pósfai G, Plunkett G 3rd, Fehér T, Frisch D, Keil GM, et al. (2006) Emergent properties of reduced-genome *Escherichia coli*. Science 312: 1044-1046.

20. Itaya M, Tsuge K, Koizumi M, Fujita K (2005) Combining two genomes in one cell: stable cloning of the Synechocystis PCC6803 genome in the *Bacillus subtilis* 168 genome. Proc Natl Acad Sci U S A 102: 15971-15976.

21. Carr PA, Church GM (2009) Genome engineering. Nat Biotechnol 27: 1151-1162.

22. Nicolaou SA, Gaida SM, Papoutsakis ET (2010) A comparative view of metabolite and substrate stress and tolerance in microbial bioprocessing: From biofuels and chemicals, to biocatalysis and bioremediation. Metab Eng 12: 307-331.

23. Brynildsen MP, Liao JC (2009) An integrated network approach identifies the isobutanol response network of *Escherichia coli*. Mol Syst Biol 5: 277.

24. Kacser H, Burns JA (1973) The control of flux. Symp Soc Exp Biol 27: 65-104.

25. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO (2004) Integrating high-throughput and computational data elucidates bacterial networks. Nature 429: 92-96.

26. Ranganathan S, Suthers PF, Maranas CD (2010) OptForce: An optimization procedure for identifying all genetic manipulations leading to targeted overproductions. PLoS Comput Biol 6: e1000744.

27. Burgard AP, Pharkya P, Maranas CD (2003) Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnol Bioeng 84: 647-657.

28. Pharkya P, Maranas CD (2006) An optimization framework for identifying reaction activation/inhibition or elimination candidates for overproduction in microbial systems. Metab Eng 8: 1-13.

29. Kim J, Reed JL (2010) OptORF: Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains. BMC Syst Biol 4: 53.

30. Park JM, Park HM, Kim WJ, Kim HU, Kim TY, et al. (2012) Flux variability scanning based on enforced objective flux for identifying gene amplification targets. BMC Syst Biol 6: 106.

31. Choi HS, Lee SY, Kim TY, Woo HM (2010) *In silico* identification of gene amplification targets for improvement of lycopene production. Appl Environ Microbiol 76: 3097-3105.

32. Yim H, Haselbeck R, Niu W, Pujol-Baxley C, Burgard A, et al. (2011) Metabolic engineering of *Escherichia coli* for direct production of 1,4-butanediol. Nat Chem Biol 7: 445-452.

33. Hua Q, Joyce AR, Fong SS, Palsson BØ (2006) Metabolic analysis of adaptive evolution for *in silico*-designed lactate-producing strains. Biotechnol Bioeng 95: 992-1002.

34. Mäder U, Schmeisky AG, Flórez LA, Stülke J (2012) SubtiWiki--A comprehensive community resource for the model organism *Bacillus subtilis*. Nucleic Acids Res 40: D1278-D1287.

35. Beisel CL, Storz G (2010) Base pairing small RNAs and their roles in global regulatory networks. FEMS Microbiol Rev 34: 866-882.

36. Storz G, Vogel J, Wassarman KM (2011) Regulation by small RNAs in bacteria: expanding frontiers. Mol Cell 43: 880-891.