Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Chemical Biology

The emergence of commodity-scale genetic manipulation

Andrea L Halweg-Edwards[1], William C Grau[2], James D Winkler[1], Andrew D Garst[1] and Ryan T Gill[1]

Since the 1970s technological advancements in the fields of synthetic biology and metabolic engineering have led to a dramatic reduction in both time and cost required for generating genomic mutations in a variety of organisms. The union of genomic editing machinery, DNA inkjet printers, and bioinformatics algorithms allows engineers to design a library of thousands of unique oligos as well as build and test these designs on a ~2 months time-scale and at a cost of roughly ~0.3 cents per base pair. The implications of these capabilities for a variety of fields are far-reaching, with potential impacts in defense, agricultural, human health, and environmental research. The explosion of synthetic biology applications over the past two decades have led many to draw parallels between biological engineering and the computer sciences. In this review, we highlight some important parallels between these fields and emphasize the importance of engineering design strategies.

Addresses
[1] Department of Chemical and Biological Engineering, University of Colorado Boulder, United States
[2] Department of Chemistry and Biochemistry, University of Colorado, United States

Corresponding author: Gill, Ryan T (rtg@colorado.edu)

Current Opinion in Chemical Biology 2015, **28**:150–155

This review comes from a themed issue on **Synthetic biology**

Edited by **Michelle CY Chang** and **Huimin Zhao**

http://dx.doi.org/10.1016/j.cbpa.2015.07.009

1367-5931/Published by Elsevier Ltd.

Introduction

The advent of high-speed computation has dramatically changed the way we approach mathematical problems [1]. For example, given a large enough dataset, recently described machine learning meta-algorithms, such as bagging (bootstrap aggregation — a method used to decrease variance of model predictions by averaging misclassification errors) and boosting (used to decrease bias and variance by weighting classifiers based on accuracy) can calculate models with arbitrarily low prediction error by starting from many (thousands of) models that perform only slightly better than random guessing [2]. Similarly, techniques that allow DNA sequence manipulation have revolutionized biology, allowing thousands of DNA oligomers to be designed, synthesized, and tested with very rapid turnaround times and at increasingly lower cost [3,4]. The ability to construct and reprogram genomes has captured the imagination, with comparisons being drawn between the current state of 'synthetic biology' and the formative days of computer science.

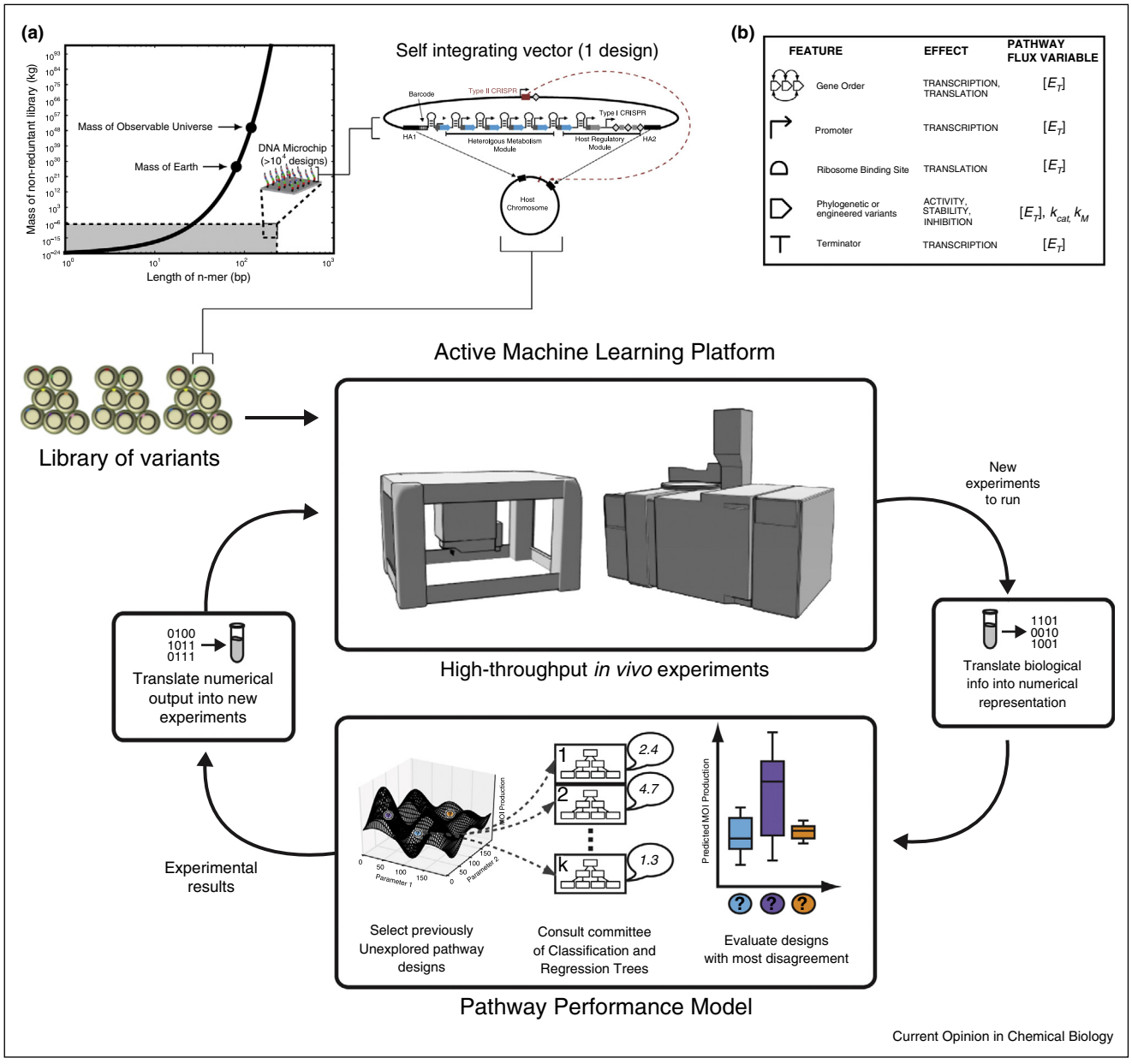
Alongside the development of high-speed computation in the 1970s, biologists began manipulating the genetic code using recombinant DNA technologies [5]. During the same period, Khorana *et al.* pioneered the *de novo* synthesis of DNA and performed the first complete construction of a gene [6,7], the yeast alanine tRNA. Ten years later, advances made by Letsinger and Caruthers allowed the rapid scaling of oligonucleotide synthesis [8–10]. Continued optimization of these techniques over the past thirty years has enabled the synthesis of gene-sized double stranded (ds)DNA fragments (<10 kbp) in as little as 4 days and oligonucleotide libraries (55 000 unique 200-mers) at costs of ~0.3 cents per base pair with turnaround times of ~1 month [11–13]. The improved economics of DNA synthesis has made it possible for biologists to build and edit genomes, thereby setting the stage for reprogramming genomes in a predictable way.

Computer scientists develop applied statistics algorithms, commonly referred to as 'machine learning' algorithms, in order to refine mathematical models that predict an outcome given a set of circumstances (such as which movies a patron will enjoy given the films that the customer has already viewed). Initially, the models offer poor predictions of outcome, often performing slightly better than random chance, and these models are referred to as 'weak hypotheses' or 'weak learners'. Following model refinement, predictions become arbitrarily well-correlated with the real classification, and are said to be 'strong hypotheses' or 'strong learners' [2]. Analogously, genetic variants can be thought of as 'weak hypotheses' if the encoded enzymes or pathways perform only slightly better than (or equivalent to) the wild-type sequence. By coupling high-throughput DNA synthesis technology with systems biology data collection (proteomics, metabolomics, transcriptomics, and DNA sequencing), engineers are now beginning to approach biological problems in a manner analogous to machine learning algorithms developed to predict social or consumer behavior patterns (Figure 1). By designing and

constructing thousands of sequences and building a database of mutational effects using high-throughput data collection, engineers can begin to generate machine learning algorithms that allow recursion through the design-build-test cycle and optimization of genetic variants.

Importantly, the ability to script and process data, does not endow a programmer with skills necessary to be a computer scientist. Rather it is the merging of statistics, mathematics, and algorithmic thinking that leads to breakthrough technologies such as the MapReduce

Figure 1



Exploration of sequence space is limited by the physical nature of the nucleic acid polymer. **(a)** Engineers are limited to a small region of sequence space highlighted by the 'DNA microchip', a cartoon illustration of a high-fidelity DNA array used to produce 1–10 pmol of oligonucleotides (up to 55 000 unique 200-mers). These designs can then be inserted into an appropriate vector backbone, transformed into a target microbe, and optionally inserted into the target genome (depending on the design strategy). The library of genetic variants can then be tested using a combination of high-throughput systems biology techniques (metabolomics, proteomics, transcriptomics, and DNA sequencing) and the data collected can be used for development of active machine learning algorithms. In this schematic, Classification And Regression Tree (CART) analysis is used to determine the combination of pathway features that result in an optimal construct. **(b)** Although DNA synthesis is limited, considerable advances can be made toward the optimization of industrial production strains by focusing on the thousands of tiny regions throughout the genome that impart large changes on flux through metabolic pathways including promoters, ribosome binding sites, enhancers, enzyme active sites, structurally stabilizing protein motifs, and allosteric regulatory domains.