Metabolic Engineering Communications 3 (2016) 227–233

Contents lists available at ScienceDirect

Metabolic Engineering Communications

journal homepage: www.elsevier.com/locate/mec

Quantifying complexity in metabolic engineering using the LASER database

James D. Winkler [1], Andrea L. Halweg-Edwards, Ryan T. Gill *

Department of Chemical and Biological Engineering, University of Colorado-Boulder, Jennie Smoly Caruthers Biotechnology Building, Research Park, Boulder, CO 80303, USA

ARTICLE INFO

Article history:
Received 18 April 2016
Accepted 4 July 2016
Available online 7 July 2016

Keywords:
Metabolic engineering
Synthetic biology
Standardization
Design tools

ABSTRACT

We previously introduced the LASER database (Learning Assisted Strain EngineeRing, https://bitbucket.org/jdwinkler/laser_release) (Winkler et al. 2015) to serve as a platform for understanding past and present metabolic engineering practices. Over the past year, LASER has been expanded by 50% to include over 600 engineered strains from 450 papers, including their growth conditions, genetic modifications, and other information in an easily searchable format. Here, we present the results of our efforts to use LASER as a means for defining the complexity of a metabolic engineering "design". We evaluate two complexity metrics based on the concepts of construction difficulty and novelty. No correlation is observed between expected product yield and complexity, allowing minimization of complexity without a performance trade-off. We envision the use of such complexity metrics to filter and prioritize designs prior to implementation of metabolic engineering efforts, thereby potentially reducing the time, labor, and expenses of large-scale projects. Possible future developments based on an expanding LASER database are then discussed.

1. Introduction

The core of engineering is the development, modification, and maintenance of complex systems to satisfy design constraints and achieve desired system behavior. Metabolic engineers, in particular, focus on producing energy, medicine, and chemical feedstocks using engineered biocatalysts (Jullesson et al., 2015). Abstracting this complexity away to simplify common design activities through the development of new methodologies, empirical design rules, and computational design tools has been a significant contributor to more effective engineering practices in a range of disciplines. Metabolic engineering has begun a similar transition into a more standardized field with the development of effective metabolic modeling techniques (King et al., 2015; Long et al., 2015), improved experimental tools (Pines et al., 2015), as well as nascent standards to disseminate strain designs (Hucka et al., 2003; Galdzicki et al., 2014; Woodruff et al., 2013). The push towards standardization will be a key focus in metabolic engineering and synthetic biology for the foreseeable future, especially as strain engineering capabilities continue to improve (Dietrich et al., 2010; Rogers and Church, 2016).

Development of these modeling and analytical tools requires data sources that provide insight into the motivations, methods, and genetic targets of a variety of engineered strains. The introduction of the LASER database (Winkler et al., 2015), containing hundreds of curated metabolic engineering designs along with associated analysis tools, provides one of the first platforms amenable to field-wide investigation. In combination with a range of other data sources that support metabolic engineering efforts (Caspi et al., 2014; McCloskey et al., 2013), LASER can provide a foundation for rigorous analysis of current metabolic engineering practices. Analogous to the efforts of software engineers to understand what makes large software projects maintainable (Weyuker, 1988), one of the first steps towards developing a standard process for strain design is to develop a definition of design complexity, as defined by metabolic engineers, that relates to metrics of design success, such as yield, titer, productivity, and ease/cost of implementation. Summarizing the difficulty of implementation and optimization of a design into a single value has allowed software engineers to analyze programs and reduce complexity as needed (Zuse, 1991); metabolic engineers may be able to leverage a similar process to quickly build biocatalysts that perform more predictably under production conditions. This need is becoming more pressing as our ability to construct strains continues to exponentially expand (Jakočiūnas et al., 2015; Zalatan et al., 2015; Horwitz et al., 2015), necessitating the development of

* Corresponding author.
E-mail addresses: james.winkler@gmail.com (J.D. Winkler), andrea.edwards@colorado.edu (A.L. Halweg-Edwards), rtg@colorado.edu (R.T. Gill).
[1] Present address: Shell Biodomain, 3333 Texas 6, Houston, TX 77082, United States.

http://dx.doi.org/10.1016/j.meteno.2016.07.002

a filtering mechanism to avoid screening of more complex designs that may not perform better than simpler alternatives.

In this study, we utilize the expanded LASER database to develop methods for evaluating the complexity of metabolic engineering designs. The Winkler-Gill complexity (WGC) metric, which estimates complexity in terms of the number and variety of mutations and techniques used to construct a design, and Frequency complexity (FC), which estimates complexity from the frequency at which mutations and methods are used in LASER designs. In order to demonstrate the applicability of WGC to common metabolic engineering problems, WGC is subsequently applied to aid in the design and filtering of libraries of distinct target complexities. Finally, we conclude by examining possible future applications for these complexity metrics in the context of increasing standardization in the metabolic engineering field.

2. Methods and materials

2.1. Data sources

The LASER database (https://bitbucket.org/jdwinkler/laser_release) contains 622 curated metabolic engineering designs, where the growth conditions, product of interest, yield and titer, and genetic modifications to the strain are recorded. Each LASER paper is associated with a perceived complexity score (1–6 scale, 6 being the most complex), based principally on the authors description of their designs and approach in the paper abstract, during the curation process. LASER records also include information regarding engineering methodology and the intent associated with particular mutations, among other pertinent design aspects (Winkler et al., 2015). Metabolic models were generated from LASER records using a combination of Biocyc backing databases (Caspi et al., 2014) to associate genes with reactions and cobrapy for metabolic model manipulation (Ebrahim et al., 2013). Regulatory models were generated directly from the published *Escherichia coli* and *Saccharomyces cerevisiae* networks (Salgado et al., 2013; Teixeira et al., 2013). All software was implemented in Python 2.7 unless stated otherwise. Statistical and numerical analysis was performed using the scipy 0.16.0b1 and numpy 1.9.2 packages.

2.2. Network representation

Both metabolic and regulatory networks are represented as a directed graph G(N,E), where N and E represent sets of vertices and edges connecting them, respectively. While regulatory networks are typically provided as directed networks, in this case, we generated metabolic networks directly from the corresponding metabolic model after excluding well-connected currency metabolites (Guimera and Amaral, 2005; Ravasz et al., 2002). A bipartite network consisting of metabolite-reaction links was used for metabolic network analysis (Jeong et al., 2000). The Python NetworkX 1.9.1 and igraph 0.7.1 packages were used for network-related computations, such as centrality, clustering, and visualization.

2.3. Synthetic library generation and analysis

For examination of high-complexity design filtering, the random library of mutated regulators was generated by identifying the 20 regulators with the highest out-degree (i.e. regulate the most proteins) in the *E. coli* and *S. cerevisiae* transcriptional networks. Once identified, every possible pairwise regulator combination was generated, and the Winkler-Gill complexity score generated by calculating the number of genes and regulatory clusters affected by simultaneous modification of both proteins. The type of mutation applied and their intent were assumed to be the same for calculating design complexity. The resulting heatmap is the calculated complexity due to genetic interactions arising from simultaneous mutation of both regulators.

Demonstration of low-complexity filtering used designs provided by Yang et al. (2011). The provided designs identify reactions that are deleted or have their minimum or maximum flux bounds altered; these specifications were converted into LASER designs by assuming each reaction alteration is due to mutation of a single gene meant to increase flux to succinate formation or reduce by-product formation. The complexity scores for the designs were then plotted alongside the predicted theoretical yield of each design to determine the yield-complexity correlation to enable library filtering.

3. Results and discussion

3.1. LASER updates

Since the initial release of LASER in 2015 (Winkler et al., 2015), our principal focus has been including additional metabolic engineering design data, developing improved visualization and analysis tools to understand where, how, and why researchers are creating these designs, and developing new metrics to guide experimentation. As a result of these efforts, LASER now contains 622 curated designs obtained from the metabolic engineering literature, split between 433 *E. coli* and 190 yeast (*S. cerevisiae*) strains. A total of 139 papers were added to the database, bringing the total to 450 curated metabolic engineering studies. This database update represents an approximate 50% increase in the size of LASER, both in terms of curated papers and deposited designs, compared to the initial release of the database (Winkler et al., 2015). The trend in papers per year along with mutations per design can be seen in Fig. 1A.

3.2. Complexity from design volume

Software engineers developed a wide-range of topological (program structure) and volume (program content) derived complexity metrics for various purposes (Zuse, 1991), mainly to limit difficulties in long-term maintenance and reduce the number and impact of errors on the desired functionality. The driving force behind these questions is the practical importance of a complexity metric and how it translates into quantifiable changes in software design, maintainability, and reusability in the future (Weyuker, 1988). Since there is no universally agreed-upon definition of complexity (Liu and Li, 2012), the most critical property of a LASER guided complexity metric is that it conforms to the perception of difficulty held by the metabolic engineering community. Volume-based metrics based on easily measurable code properties such as operator use, such as the original Halstead metric (Halstead, 1975), have been employed by software engineers for decades in an attempt to identify complexity within programming projects. Volume metrics are particularly tractable for adaptation by metabolic engineers, as they measure values with readily identifiable analogs in most designs: the number of genes mutated in an organism (η_1), the variety of methods used to introduce these mutations (η_2)), how the manipulated components of the metabolic and regulatory networks interact (η_3), and the intended effect of each mutation (η_4).

The principal challenge then is to convert these properties into a score that describes . One reasonable way to assess complexity of building one strain is to calculate the expected number of effects per gene modification (Eq. (1), Fig. 1B), under the assumption that modifications that affect gene or cellular physiology more severely

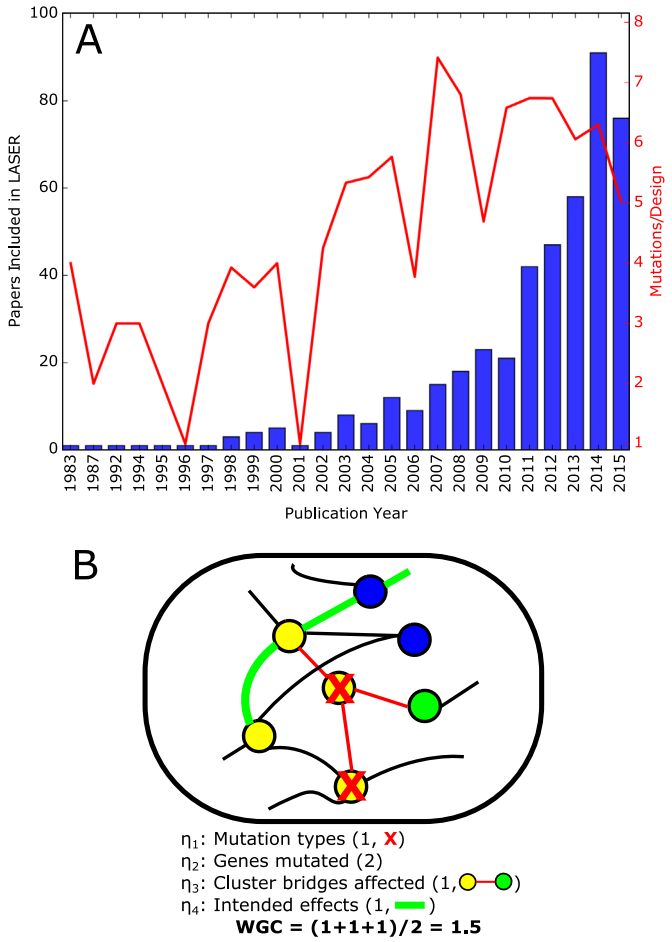

Fig. 1. (A). Trends in number of metabolic engineering papers published per year, along with the average number of mutations per strain (red line) in their designs. Years 1983–1997 and 2001 contain only a single datapoint. B). Calculation of $\eta_1 - \eta_4$, along with the WGC score, for a single mutant. Node clusters are denoted by color. In this case, there is one mutation type (deletion, X; $\eta_1 = 1$), two mutated genes ($\eta_2 = 2$), one edge between a cluster containing a modified gene and a non-modified gene ($\eta_3 = 1$), and one intended effect of increasing the flux through part of the metabolic network ($\eta_4 = 1$). The resulting WGC score is $(\eta_1 + \eta_3 + \eta_4)/\eta_2 = 1.5$. For study duration, these properties are calculated from all mutants described in the papers LASER record. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

are difficult to optimize. At the level of an entire study, it is necessary to holistically use the mutant properties to estimate the time needed to complete the entire study. One common approach to similar problems in the social sciences is to use multilinear regression to use a collection of independent variables to predict some dependent outcome of interest (King, 1986); rather than explicitly imitating the form of the original Halstead equation, we instead represent complexity as a linear function of these variables computed from all mutants described in a LASER record. (Eq. (2)). Standard regression procedures can be used to determine the values of the constants α_i if the time required to complete a subset of LASER studies is known.

$$C_{WGC} = \frac{\eta_1 + \eta_3 + \eta_4}{\eta_4}(1 + \eta_5/\eta_2) \quad (1)$$

$$T_{WGC} = \sum_i \alpha_i \sum_j^N \eta_{i,j} \quad (2)$$

3.3. Complexity from modification frequency

Complexity is not only conceptualized as a function of how challenging it is to implement a design. Well-trodden experimental lines of thought, though they at times may be involved and require extensive strain re-engineering and long periods of time to complete, have the backing of experience from the field and prior technique optimization to simplify the engineering process. Leveraging new experimental approaches lacking decades of widespread use for metabolic engineering poses unique demands that are not capture by only examining the properties of the final strains but also how frequently a given technique is used for strain engineering over the entire LASER corpus. Intuitively, frequently used modifications or techniques should be considered simpler as they become increasingly refined following their introduction to the field. Defining a complexity metric based on how frequently particular modifications and approaches to metabolic engineering problems are used in the LASER dataset (C_f, Eq. (3)) is quite straightforward. The frequency of the ith mutation ($f_{mut,i}$) and jth method ($f_{met,j}$) are computed by dividing the count of each mutation or method by the total number of mutations or methods (respectively) used in the entire LASER database. The complexity of a given design is then calculated by calculate the sum of inverse frequencies for both sets, on the hypothesis that less frequently employed strain engineering approaches are complex compared to commonly employed methods or mutation types.

$$C_f = \sum_i^{N_{mut}} 1/f_{mut,i} + \sum_j^{N_{met}} 1/f_{met,j} \quad (3)$$

One property of this metric that distinguishes it from the volume-based WGC approach is that C_f will decrease for each over time as additional designs are deposited into LASER, rather than being an immutable product of the design genotype and design approach. New ways of controlling gene function and methodologies for introducing mutations will form the “bleeding” complexity edge in the dataset, which will then decrease in complexity over time. It should be an effective adjunct to the more effort-focused WGC metric when estimating the complexity of implementation or the anticipated research impact of the design on the metabolic engineering field.

3.4. Volume-based analysis

With two candidate definitions of design complexity in hand, the entire LASER dataset can be analyzed to determine trends in complexity over time and by product classification. In the case of WGC, the distribution of complexity (Fig. 2A) is highly skewed towards the low complexity ends, due to existence of many designs with few mutations and limited interaction with native regulatory and metabolic networks. The papers containing the most complex designs generally use techniques that can result in gross modifications of large genomic regions (Utrilla et al., 2012) or involve wide-ranging modification of multiple metabolic and regulatory modules (Santos et al., 2012; Raman et al., 2014). In general, designs that require many effects and mutation types, or interact with a large number of clusters in host regulatory and metabolic networks should have the highest complexity scores.

Complexity, as measured by WGC, does not predict design performance: WGC is not significantly correlated with product yields (P=0.08, Spearman) for the 109 designs that include this information. The lack of an observable relationship between these properties suggests that yield improvements are generally incidental to academic metabolic engineering projects, which may instead be focused on validating new technologies, pathways, or particular mutational targets. There is a very weak correlation

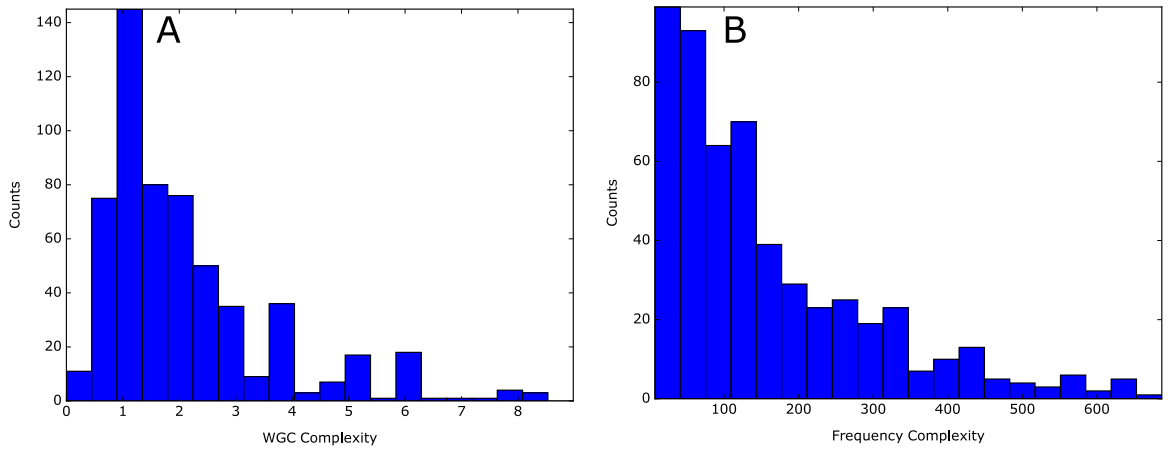

Fig. 2. The complexity distribution for (A) WGC and (B) frequency (C_f) metrics for the entire LASER dataset. Values greater than five times the median complexity are placed in the final bin for each histogram. Median WGC complexity is 1.79, while median frequency complexity is 137.4; both distributions indicate LASER is highly skewed towards low complexity designs.

between WGC and the perceived complexity (as recorded by the original LASER record curator, see Methods and Materials) of each study ($P=0.025$, $r=0.09$), which follows the idea that experimenters rarely judge the complexity of a design or paper using the raw number or type of mutations used for engineering. WGC may be taken to represent the diversity of approaches required to modify a strain to perform as needed, so that strains with high WGC are more likely to require significant efforts to optimize. In the absence of a large experimental library of strains to test this hypothesis, we can apply this metric to sets of strains generated through computational means to determine which strains will presumably require more effort to construct.

3.5. Frequency-based trends

Unlike WGC, C_f has no dependence on the underlying metabolic or regulatory topology of an organism; the complexity of a design is derived purely from the frequency that the mutation types and methods are used in the LASER database. Due to the design of the metric, we also observe a similarly skewed distribution towards the bulk of studies using well-worn metabolic engineering approaches, with a leading edge of projects introducing new approaches to strain design (Fig. 2B). One possible conceptualization of this metric is that it attempts to capture the difficulty of utilizing new, less mature approaches to strain engineering than classical methods, while WGC is more focused on describing the engineering complexity of the desired strain design. However, there is still little correlation between complexity and yield (P=0.81, Spearman), again indicating that additional complexity, as measured using both of these metrics, does not necessarily translate into better design performance in academic research. We expect a significantly different complexity-yield tradeoff for industrial designs, but we cannot test this hypothesis due to the paucity of publicly available data.

Given that C_f is explicitly designed to quantify the concept of novelty, it is encouraging that we observe a significant and relatively large correlation between perceived complexity and C_f ($P < 10^{-11}$, $r=0.28$). The presumable key to being perceived as highly complex is the use of innovative new techniques to accomplish challenging metabolic engineering goals, which is the essential core of the frequency complexity metric. Interestingly, there is also a significant positive correlation between C_f and the estimated time required to complete a study ($P = 4.2{*}10^{-14}$, see below), implying that additional time is needed on account of the novel techniques employed. Overall, frequency complexity efficiently describes the novelty of designs, allowing researchers to quickly estimate if they are on well-trodden ground in their research.

3.6. Harnessing complexity for experimental design

Due to the advent of increasingly inexpensive DNA synthesis and sequencing technology, it is now possible to routinely generate genome-scale libraries that modify many genomic loci or saturate a small number of sites with mutations. However, the cost of these DNA libraries remains high, as do many subsequent screening strategies, and so the selection of optimal targets for mutagenesis remains critical for achieving the desired phenotypic outcomes. The complexity metrics developed here provided a metric for assessing how physiologically disruptive particular combinations of mutations are, allowing researchers to estimate the impact of the library on the host strain. This ability is particularly useful when manipulating global regulatory genes by enabling the identification of genotypes that lead to the maximum amount of wide-ranging gene expression changes so that if a suitable selection is available, evolutionary optimization can be applied to identify strains with the desired production or tolerance phenotype. A similar procedure can also be used to examining variants meant to achieve a common phenotypic goal, such as succinic acid production.

3.7. Library design

In order to test how our complexity metrics functioned for design discrimination, a synthetic random library of strains was created by mutating (in silico) pairs of regulators in *E. coli* and *S. cerevisiae* that regulate the largest number of genes in each organism. This procedure is analogous to making a random library meant to identify phenotypes arising from mutating global regulators, and has proven to be an effective strategy for improving both tolerance and production phenotypes in the past (Santos et al., 2012; Huang et al., 2015). A total of 20 regulators in each organism were targeted for in silico mutation, and the WGC score for each design in the synthetic library was computed. In the case of *E. coli*, many pairs of regulatory mutations that have wide-ranging effects on gene expression, especially those involving RpoD mutations. Mutations in this particular sigma factor have been used in the past for enhancing L-tyrosine yield (Santos et al., 2012) due to its role in the expression of genes required for vegetative growth. Other potential high-complexity mutation pairs involving other sigma factors, CRP (Geng and Jiang, 2015), and FNR

could be candidates for more thorough mutagenesis. Although these results are intuitive based on current knowledge of the respective regulatory networks in each organism, the addition of WGC permits quantifying the level of transcriptional disruption associated with each mutational combination so that the appropriate perturbations can be selected.

3.8. Design properties to study duration

In order to estimate how the properties measured by WGC translated into the amount of time needed to complete a given study, we contacted metabolic engineering groups whose strains had been curated in LASER to determine the amount of time required from conceptualization to initial submission of the publication and obtained 19 data points. Since the estimates were highly sensitive to the number of clusters affected by strain manipulations (η_3, we applied a binary filter that converted η_3 to zero if no cross-cluster edges were affected, and incremented by per mutant otherwise. We also included another variable η_5 in the regression to represent the total library or strain set size used, since time to completion should depend on the number of strains that must be built. In order to estimate full-time effort from the provided data, we estimated the contribution of each author using the sum of a harmonic series out to $N_{authors}$ to account for the presumably decreasing effort on the part of contributors beyond the first author. Detailed analyses of author-effort correlations are not available, but this approximation should help to account the benefits and costs of collaboration among increasingly large working groups.

Linear regression of the sum of η_1–η_5 properties for each paper with the provided 19 estimates yields the final correlation shown in Fig. 3A, which accounts for a large proportion of the observed variance in full-time effort required for study completion. ($R^2 = 0.77$). Applying the correlation to the LASER dataset reveals that the median study requires approximately 4 years of effort to complete (Fig. 3B, with a maximum of 10 years effort. To our knowledge, this correlation between properties of the engineered strains constructed in each study and estimated time to study completion the first of its kind for metabolic engineering, and should be able to provide general guidance to researchers in the planning phases of their projects.

The most important question concerning these time estimates is whether the additional effort translates into a tangible improvement in the implemented designs. In this case, no significant correlation between estimated completion time and yield (P=0.18) was detected, perhaps due to the fact that yield maximization is rarely the only goal in most researcher projects. Interestingly there is a relatively strong negative correlation ($P = 3.9*10^{-25}$, r = − 0.40) between frequency complexity and time to completion, implying that more complex studies require less time to complete; this surprising finding suggests novel studies use a lower diversity of modifications compared to others in order to validate new technologies. The simplest explanation for this phenomenon is that new techniques require significant effort to validate, so it is conceivable that researchers will choose the minimal test case required for their new approach. More

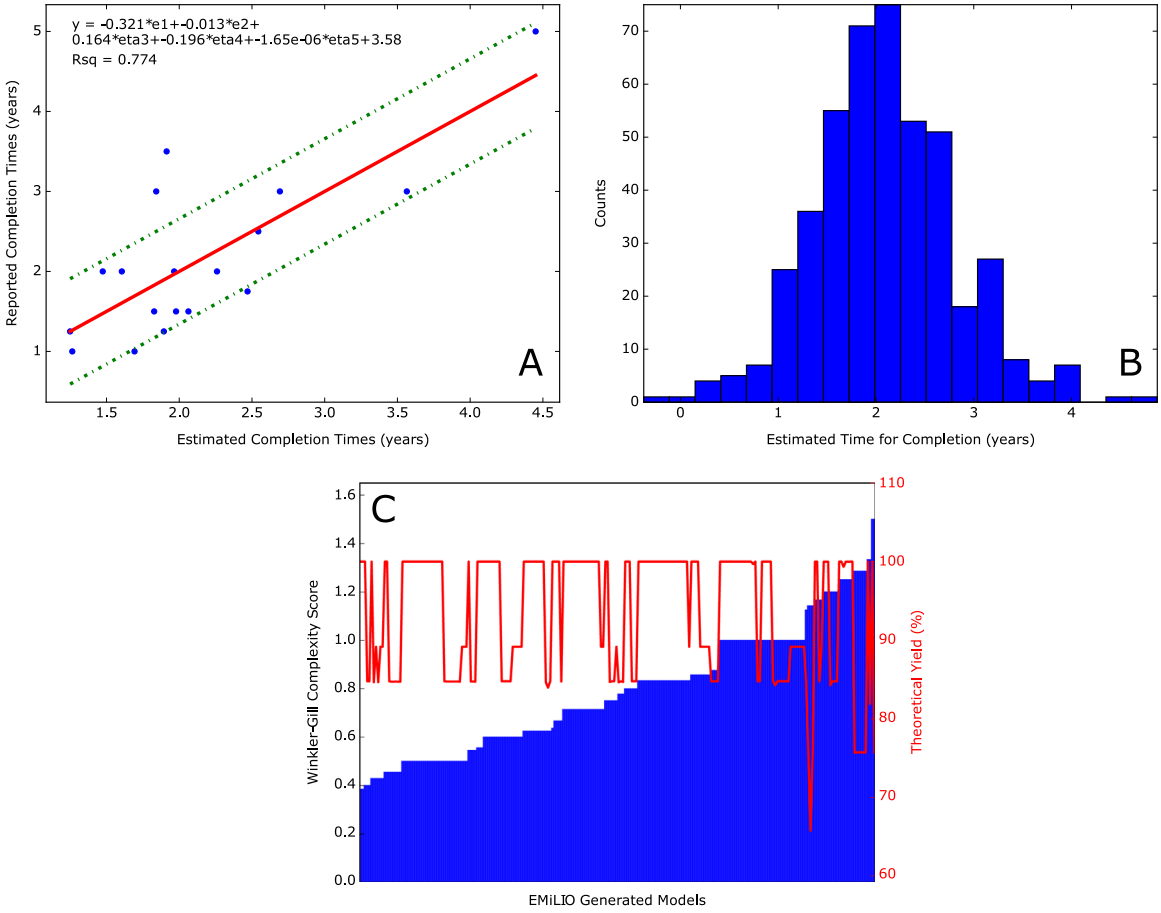

Fig. 3. The (A) correlation between LASER-extracted topological properties and experimenter-reported study lengths (dashed lines denoted the 90% confidence interval for estimates), and (B) distribution of LASER study time estimates generated using the correlation. Multilinear regression was performed using the scipy optimize package using Eq. (2). C). The Winkler-Gill design complexities of *E. coli* proposed succinic acid production strains, along with their predicted theoretical yields from glucose (right-hand y-axis). Each proposed design was converted into a LASER design as discussed in Methods and Materials and analyzed using the same analysis pipeline.

traditional approaches can be applied on a larger-scale due to prior method derisking. If greater amounts of data were available concerning researcher effort, funding, and the probability of unexpected delaying events known, a superior correlation accounting for these factors as well could be constructed to provide improved results.

3.9. Library filtering

On the other end of the complexity spectrum, complexity metrics can be applied to proposed strains identified by computational strain design algorithms to identify the least complex designs achieving the desired performance (yield). In one such study, the EMiLIO algorithm was used to generate 234 designs with enhanced succinate yield under aerobic and anaerobic conditions, compared to wild-type *E. coli* (Yang et al., 2011). Complexity arises both from the number of reactions to be manipulated and their effect on multiple metabolic clusters within the *E. coli* metabolic network; since the method for implementing these mutations was not specified, the complexity scores here represent the minimal complexity possible when implementing the prescribed models. In this case, low-complexity designs usually have fewer mutations.

Analyzing the complexity of proposed models reveals the expected lack of correlation between WGC score and the corresponding predicted theoretical yield from EMiLIO (Fig. 3C). Given that the low complexity designs are predicted to achieve approximately the same theoretical yield as those at the high complexity end of the model distribution, a large proportion of the 234 designs can be eliminated outright from the pool of designs for experimental implementation without impacting the maximum achievable succinate yields. This filtering step has obvious advantages for individual laboratories and larger-scale organism foundries, as it reduces both the number of strains and required number of mutations to achieve the desired yield. This approach works best when evaluating a collection of designs with different combinations of gene mutations; in cases where a set of genes are modified in different ways, such as site-directed mutagenesis of coding sequences, all of the models would have identical complexities. However, the vast majority of modifications in LASER remain deletion, overexpression, genomic integration, or plasmid cloning, so in practice this difficulty is negligible. Overall, the combination of the LASER analysis pipeline and existing strain design tools would enable rapid implementation of a design-filter-build cycle that more tractable with current metabolic engineering capabilities.

4. Conclusions

In this study, we present two formal complexity metrics for metabolic engineering designs as part of an effort quantify current metabolic engineering practices using the LASER database. The expanded database has been improved significantly since its initial publication, and now contains 622 strain designs for *E. coli* and yeast, along with myriad software improvements. We have used these data to develop two distinct ways of assessing the complexity associated with metabolic engineering designs: the Winkler-Gill complexity metric captures the effort required to actually generate a given design, while the frequency complexity metric measures the novelty of a particular design compared to the rest of the metabolic engineering field. As metabolic engineering practices become more formalized and process engineering data more readily available, it will be possible to directly correlate design complexity with expected cost of research and implementation. Even with the currently limited dataset, we were able to estimate the time required to complete all LASER studies by correlating the properties measured by WGC with time estimates provided from practicing metabolic engineers; time and data will only improve the accuracy of these metrics and correlations, as was the case for other engineering fields.

More immediately, we expect complexity metrics to play an important role in the filtering of potential designs generated by large-scale organism foundries to avoid unnecessary strain construction and screening. This sieving process, demonstrated here using both random libraries and computationally generated strain designs, would enable significantly higher throughput through the foundry until advances in supporting technologies have bridged the gap between construction and evaluation capabilities. As the desired scale of organism engineering becomes ever larger, complexity-based approaches for reducing experimental workload will become key parts of the metabolic engineering design cycle.

Acknowledgments

We thank the Department of Energy Genome Science Program (award #DE-SCOOO8812) for funding and the Hsueh-Fen Juan laboratory for helpful conversations during the course of writing this manuscript.

References

Caspi, R., Altman, T., Dreher, K., Fulcher, C.A., Subhraveti, P., Keseler, I.M., Kothari, A., Krummenacker, M., Latendresse, M., Mueller, L.A., Ong, Q., Paley, S., Pujar, A., Shearer, A.G., Travers, M., Weerasinghe, D., Zhang, P., Karp, P.D., 2014. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 42 (D1), D459–D471.

Dietrich, J.A., McKee, A.E., Keasling, J.D., 2010. High-throughput metabolic engineering: advances in small-molecule screening and selection. Annu. Rev. Biochem. 79, 563–590.

Ebrahim, A., Lerman, J.A., Palsson, B.O., Hyduke, D.R., 2013. Cobrapy: constraints-based reconstruction and analysis for Python. BMC Syst. Biol. 7 (1), 74.

Galdzicki, M., Clancy, K.P., Oberortner, E., Pocock, M., Quinn, J.Y., Rodriguez, C.A., Roehner, N., Wilson, M.L., Adam, L., Anderson, J.C., Bartley, B.A., Beal, J., Chandran, D., Chen, J., Densmore, D., Endy, D., Grünberg, R., Hallinan, J., Hillson, N.J., Johnson, J.D., Kuchinsky, A., Lux, M., Misirli, G., Peccoud, J., Plahar, H.A., Sirin, E., Stan, G.-B., Villalobos, A., Wipat, A., Gennari, J.H., Myers, C.J., Sauro, H.M., 2014. The synthetic biology open language (SBOL) provides a community standard for communicating designs in synthetic biology. Nat. Biotechnol. 32 (6), 545–550.

Geng, H., Jiang, R., 2015. Camp receptor protein (crp)-mediated resistance/tolerance in bacteria: mechanism and utilization in biotechnology. Appl. Microbiol. Biotechnol., 1–11.

Guimera, R., Amaral, L.A.N., 2005. Functional cartography of complex metabolic networks. Nature 433 (7028), 895–900.

Halstead, M.H., 1975. Toward a theoretical basis for estimating programming effort. In: Proceedings of the 1975 annual conference, ACM, pp. 222–224.

Horwitz, A.A., Walter, J.M., Schubert, M.G., Kung, S.H., Hawkins, K., Platt, D.M., Hernday, A.D., Mahatdejkul-Meadows, T., Szeto, W., Chandran, S.S., Newman, J. D., 2015. Efficient multiplexed integration of synergistic alleles and metabolic pathways in yeasts via CRISPR-Cas. Cell Systems.

Huang, L., Pu, Y., Yang, X., Zhu, X., Cai, J., Xu, Z., 2015. Engineering of global regulator camp receptor protein (crp) in *Escherichia coli* for improved lycopene production. J. Biotechnol. 199, 55–61.

Hucka, M., Finney, A., Sauro, H.M., Bolouri, H., Doyle, J.C., Kitano, H., the rest of the SBML forum:, Arkin, A.P., Bornstein, B.J., Bray, D., Cornish-Bowden, A., Cuellar, A. A., Dronov, S., Gilles, E.D., Ginkel, M., Gor, V., Goryanin, I.I., Hedley, W.J., Hodgman, T.C., Hofmeyr, J.-H., Hunter, P.J., Juty, N.S., Kasberger, J.L., Kremling, A., Kummer, U. , Le Novre, N., Loew, L.M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E.D., Nakayama, Y., Nelson, M.R., Nielsen, P.F., Sakurada, T., Schaff, J.C., Shapiro, B.E., Shimizu, T.S., Spence, H.D., Stelling J., Takahashi, K., Tomita, M. Wagner, J., Wang, J., 2003. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 19 (4), 524–531.

Jakočiūnas, T., Bonde, I., Herrgård, M., Harrison, S.J., Kristensen, M., Pedersen, L.E., Jensen, M.K., Keasling, J.D., 2015. Multiplex metabolic pathway engineering using CRISPR/Cas9 in *Saccharomyces cerevisiae*. Metab. Eng. 28, 213–222.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., Barabási, A.-L., 2000. The large-scale organization of metabolic networks. Nature 407 (6804), 651–654.

Jullesson, D., David, F., Pfleger, B., Nielsen, J., 2015. Impact of synthetic biology and metabolic engineering on industrial production of fine chemicals. Biotechnology Advances.

King, G., 1986. How not to lie with statistics: avoiding common mistakes in quantitative political science. Am. J. Political Sci., 666–687.
King, Z.A., Lloyd, C.J., Feist, A.M., Palsson, B.O., 2015. Next-generation genome-scale models for metabolic engineering. Curr. Opin. Biotechnol. 35, 23–29.
Liu, P., Li, Z., 2012. Task complexity: A review and conceptualization framework. Int. J. Ind. Ergon. 42 (6), 553–568.
Long, M.R., Ong, W.K., Reed, J.L., 2015. Computational methods in metabolic engineering for strain design. Curr. Opin. Biotechnol. 34, 135–141.
McCloskey, D., Palsson, B.O., Feist, A.M., 2013. Basic and applied uses of genomescalemetabolic network reconstructions of *Escherichia coli*. Mol. Syst. Biol. 9 (1).
Pines, G., Freed, E.F., Winkler, J.D., Gill, R.T., 2015. Bacterial recombineering – genome engineering via phage-based homologous recombination, ACS Synthetic Biology.
Raman, S., Rogers, J.K., Taylor, N.D., Church, G.M., 2014. Evolution-guided optimization of biosynthetic pathways. Proc. Natl. Acad. Sci. 111 (50), 17803–17808.
Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N., Barabási, A.-L., 2002. Hierarchical organization of modularity in metabolic networks. Science 297 (5586), 1551–1555.
Rogers, J.K., Church, G.M., 2016. Multiplexed engineering in biology. Trends Biotechnol. 34 (3), 198–206.
Salgado, H., Peralta-Gil, M., Gama-Castro, S., Santos-Zavaleta, A., Muniz-Rascado, L., Garcia-Sotelo, J.S., Weiss, V., Solano-Lira, H., Martinez-Flores, I., Medina-Rivera, A., Salgado-Osorio, G., Alquicira-Hernandez, S., Alquicira-Hernandez, K., Lopez-Fuentes, A., Porron-Sotelo, L., Huerta, A.M., Bonavides-Martinez, C., Balderas-Martinez, Y.I., Pannier, L., Olvera, M., Labastida, A., Jimenez-Jacinto, V., Vega-Alvarado, L., del Moral-Chavez, V., Hernandez-Alvarez, A., Morett, E., Collado-Vides, J., 2013. Regulondb v8. 0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. Nucleic Acids Res. 41 (D1), D203–D213.
Santos, C.N.S., Xiao, W., Stephanopoulos, G., 2012. Rational, combinatorial, and genomic approaches for engineering l-tyrosine production in *Escherichia coli*. Proc. Natl. Acad. Sci. 109 (34), 13538–13543.
Teixeira, M.C., Monteiro, P.T., Guerreiro, J.F., Goncalves, J.P., Mira, N.P., dos Santos, S. C., Cabrito, T.R., Palma, M., Costa, C., Francisco, A.P., Madeira, S.C., Oliveira, A.L., Freitas, A.T., Sa-Correia, I., 2013. The yeastract database: an upgraded information system for the analysis of gene and genomic transcription regulationin *Saccharomyces cerevisiae*. Nucleic Acids Res., gkt1015
Utrilla, J., Licona-Cassani, C., Marcellin, E., Gosset, G., Nielsen, L.K., Martinez, A., 2012. Engineering and adaptive evolution of *Escherichia coli* for d-lactate fermentation reveals gatc as a xylose transporter. Metab. Eng. 14 (5), 469–476.
Weyuker, E.J., 1988. Evaluating software complexity measures. IEEE Trans. Softw. Eng. 14 (9), 1357–1365.
Winkler, J.D., Halweg-Edwards, A.L., Gill, R.T., 2015. The LASER database: Formalizing design rules for metabolic engineering. Metabolic Engineering Communications.
Woodruff, L., May, B.L., Warner, J.R., Gill, R.T., 2013. Towards a metabolic engineering strain commons: an escherichia coli platform strain for ethanol production. Biotechnol. Bioeng. 110 (5), 1520–1526.
Yang, L., Cluett, W.R., Mahadevan, R., 2011. EMILIO: a fast algorithm for genome-scalestrain design. Metab. Eng. 13 (3), 272–281.
Zalatan, J.G., Lee, M.E., Almeida, R., Gilbert, L.A., Whitehead, E.H., Russa, M.L., Tsai, J. C., Weissman, J.S., Dueber, J.E., Qi, L.S., Lim, W.A., 2015. Engineering complex synthetic transcriptional programs with CRISPR RNA scaffolds. Cell 160 (1), 339–350.
Zuse, H., 1991. Software complexity, NY, USA: Walter de Cruyter.