

UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

Tools for engineering biology: methods for designing, building, and testing

Permalink

<https://escholarship.org/uc/item/1w32c0hq>

Author

Hsiao, Timothy

Publication Date

2013

Peer reviewed|Thesis/dissertation

Tools for engineering biology: methods for designing, building, and testing

By

Timothy H. Hsiau

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Joint Doctor of Philosophy
with University of California, San Francisco

in

Bioengineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Assistant Professor J. Christopher Anderson

Professor Patricia C. Babbitt

Associate Professor Kathleen R. Ryan

Assistant Professor John E. Dueber

Fall 2013

Copyright 2013

Timothy H Hsiau

All rights reserved

Abstract

Tools for engineering biology: methods for designing, building, and testing

by

Timothy H Hsiau

Doctor of Philosophy in Bioengineering

University of California, Berkeley

Assistant Professor J. Christopher Anderson, Chair

Genetic engineering remains a difficult task and the design, build, test cycle may take months or years to complete. Currently, all three aspects are laborious, expensive, and mostly handle volumes of tens of units. The typical process of reaching a proof of concept genetic prototype involves an intensive survey of literature, synthesizing or acquiring genes, and testing their function. Here I outline tools to address bottlenecks in the genetic engineering workflow. First, I describe the Engineered DNA Sequence Syntax Inspector (EDSSI). This software pipeline checks protein-coding DNAs for syntax errors, which are incorrect or missing elements in the DNA. By using EDSSI, researchers are able to avoid the simple but costly mistakes of point errors, misannotated gene structure, or unintended extraneous ORFs. Second, I describe Multiplex Ortholog Library Synthesis and Expression Testing (MOLSET), a method to build genes and test their expression in *E. coli*. MOLSET enables the multiplex synthesis and expression testing of up to a hundred genes directly from a microarray oligo pool. data generated by MOLSET is incorporated into a design synthesis algorithm called Act Synthesizer that employs this information to improve predictions of pathways. Finally, we show the usefulness of the Act synthesizer by designing and testing a pathway for product of the household painkiller, acetaminophen, in *E. coli*.

Table of Contents

Abstract		1
Table of Contents		
Chapter 1:	The Engineering DNA Syntax Sequence Inspector	1
Chapter 2:	Multiplex Ortholog Library Synthesis and Expression Testing	12
Chapter 3:	The Act Synthesizer: Design and Testing of Biosynthetic Pathways	26
Chapter 4:	Conclusions	34

Chapter 1

The Engineered DNA Sequence Syntax Inspector

Reproduced with permission from ACS Synthetic Biology, submitted for publication. Unpublished work copyright 2013 American Chemical Society.

Abstract

DNAs encoding for polypeptides often contain design errors that cause experiments to prematurely fail. One class of design errors is incorrect or missing elements in the DNA, referred to here as syntax errors. We have identified three major causes of syntax errors: point mutations from sequencing or manual data entry, gene structure misannotation, and unintended ORFs. EDSSI is an online bioinformatics pipeline that checks for syntax errors through three steps. First, ORF prediction in input DNA sequences is done by GeneMark; next, homologous sequences are retrieved by BLAST; and finally, syntax errors in the protein sequence are predicted by using the SIFT algorithm. We show that EDSSI is able to identify previously published examples of syntactical errors and also show that our indel addition to the SIFT program is 97% accurate on a test set of *E. coli* proteins. EDSSI is available at <http://andersonlab.qb3.berkeley.edu/Software/EDSSI/>

Preface

The contents of this chapter are based on a paper submitted to ACS Synthetic Biology. My contributions to this work included developing the DNA Syntax Inspector concept, writing code, and writing this chapter. Dr. J. Christopher Anderson contributed to developing the ideas presented in this chapter.

Introduction

Designed DNAs encoding for polypeptides often contain design errors that cause experiments to prematurely fail; to address this concern, we have developed a computational tool that detects likely syntactic design errors in a genetic construct. DNA sequences are typically designed from prior knowledge of biological phenomena but implementation of novel biological functions is error prone [1]. The cause of error requires some experimental “debugging” to discover, and typically only after ruling out many alternative hypotheses is the problem traced to incorrect or missing features in the designed DNA.

Previously, we have reviewed the many challenges that genetic engineers face [1]. One class of challenge is incorrect or missing elements in the DNA, referred to here as syntax errors. These errors can be predicted beforehand and corrected for in the design stage; however, in many experiments today they are usually discovered during the debugging stage after an experiment has failed.

Syntax errors that occur in polypeptide-coding DNAs result in a non-functional protein or cause unintended interactions in the early, proof-of-concept stages of a project. Such a result can then be interpreted as a total failure of the experiment rather than an artifact due to syntax errors. We have identified three major sources of syntax errors and

the corresponding manifestations: 1) sequencing errors in the primary data that lead to point mutations or truncations; 2) wrong gene structure annotations, typically of the gene start site that lead to truncated proteins; and, lastly, 3) unanticipated ORFs, which give rise to unintended polypeptides. We give examples for these three types of syntax errors and present an analysis pipeline that aids in the identification of such errors.

Point errors from sequencing or manual data entry: Sequencing errors from Sanger technology in the 1990s were estimated at 0.1% [2]. Although next generation sequencing can compensate for higher error rates in individual reads by using information from overlapping reads, finished contigs are still imperfect were were estimated to have an error rate of 0.33% in 2009 [3]. Additionally, error rates are unequally distributed across sequenced genomes and fluctuate based on both local sequence composition and the specific sequencing technology employed. Sequencing errors can cause nonsynonomous mutations and truncations of a gene by introducing erroneous start or stop codons. Additionally, manual sequence editing has the potential of introducing this and other types of errors.

Real world example: Engineers refactored a *Klebsiella* nitrogen-fixation gene cluster to remove unwanted regulation by synthesizing sequences derived from NCBI entry X13303.1 [4]. The synthesized genes were non-functional when tested by knockout complementation, and the failures were traced back to non-synonymous mutations due to erroneous sequencing data in the original submission. Identifying the problem and correcting it by resequencing the source DNA consumed 3 months [5].

Gene structure misannotation: While genome annotation has gone through a rapid pace of development, predicted gene structures are still imperfect and also many erroneous entries exist in the databases. As one example, automated gene annotation software mis-annotates at least 10% of prokaryotic gene start sites [6, 7]. Similarity-based analyses of genome sequences have identified gene-calling errors as high as 15% [8]. The gold standard for gene start site is experimental validation by N-terminal sequencing, which is sparse or not collected in a central database. While there have been efforts to improve annotation of gene start sites [9], many of the entries in nr, NCBI's non-redundant protein database, still have erroneous annotations. Gene prediction in eukaryotes, especially non-model organisms, is non-trivial as the software must also accurately predict introns [10]. Gene structure misannotation can also happen when users manually infer the incorrect gene structure.

Real world example: The *invF* gene was used in a design for genetic logic gates in *E. coli*; however, due to an incorrect annotation, the synthesized ORF was truncated. [11] This error is particularly common when refactoring overlapping ORFs or when dealing with ORFs that have many methionines near the start (e.g. beta-lactamase).

Unintended ORFs: overlapping ORFs have been found in all domains of life. On average, 27% of genes in prokaryotic genomes are involved in at least one overlap [Lillo], and internal or partial ORFs can occur when sequences are copied from their native context. During transfer of a target ORF to a new context, annotation of the overlapping ORF may be forgotten or discarded. Additionally, ORFs expressed outside

of their native context can contain unintended translational signals that lead to production of truncated protein products [13].

Real world example: A chimeric gene composed of the rabbit structural capsid protein VP60 fused to cholera toxin B subunit was not stable in *E. coli* hosts [14]. Constructing frame-shift mutations versions of the gene did not alleviate plasmid instability, and the true cause of instability was found to be due to the use of non-standard codons, which, when read in a different frame, led to expression of a leucine-rich ORF. The leucine-rich polypeptide was hypothesized to insert into the membrane and was shown to be the cause of toxicity.

Errors in designed protein coding sequences are predictable, preventable, and cause unnecessary experimental delays. Thus, there is a need for software tools that decrease risk of failure by identifying potential errors in a genetic design. In this article, we report the Engineered DNA Sequence Syntax Inspector (EDSSI), a new tool that identifies syntax errors in the user's protein-coding sequence. We focus on protein-coding syntax for genes from any source being expressed in bacteria, to facilitate the common practice of placing existing protein-coding sequences under engineered transcriptional regulation. Additionally, protein-coding syntax is much better understood than non-protein-coding syntax and poses a more tractable problem. In our tool, users input a DNA sequence, protein-coding regions are detected, and a homology-based approach is used to predict errors. Users can then view the syntax error analysis on the results page. By quickly allowing syntax errors to be considered or discarded as a hypothesis in troubleshooting experiments, this tool enables a more rational design of protein-coding sequences.

Methods

The sequence inspector workflow is illustrated in Figure 2. The workflow consists of the following steps: sequence submission, gene prediction, homolog search, alignment, scoring, and display.

Bioinformatics workflow

The sequence inspector predicts genes in input DNA by using GeneMark.hmm and NCBI's Conserved Domain search (CD-search). The HMM framework of GeneMark.hmm uses the statistical patterns of nucleotides coding for proteins to predict likely genes. Predictions of translational start sites are further improved by incorporating a model of the ribosome binding site (RBS). The CD-search identifies nucleotide regions matching protein family profiles [15]. The protein family profile match regions are then extended to the closest start and stop codon for a minimal gene prediction. The predicted genes from the two approaches are then merged if they overlap and are in the same frame.

The sequence inspector next searches for and retrieves homologous protein sequences, aligns the sequences, and scores the input sequence for syntax errors. To find protein homologs, for each predicted ORF, the program uses a BLAST search against the non-redundant (nr) protein database to find closely related genes. Full-length protein

sequences identified by the BLAST search are then retrieved, and the homologs are aligned using the MUSCLE aligner [16]. The alignment is then scored using the SIFT algorithm [17]. In brief, amino acid distributions at each column are used to calculate a normalized probability that the observed residue is correct. Aligned columns with more variation are more likely to tolerate substitutions than highly conserved positions. The standard cutoff of 0.05 was used. Because SIFT ignores gaps in the input alignment, we added a custom scoring function for the gapped positions that uses a simple weighted vote.

Results are passed as a JavaScript Object Notation (JSON) file and displayed using the JavaScript visualization library RaphaelJS and a JavaScript multiple sequence aligner viewer developed in-house. We use the ELink functionality provided by NCBI to retrieve publications relevant to each protein BLAST hit. Results from the analysis are output as an independent json file, which is read and displayed by the HTML/js viewer.

Workflow technical details

Genemark.hmm version 2.8a is run with the *E. coli* model and the -r option, which uses an RBS model for start codon prediction. All other prediction options were kept as default. Conserved Domain search was performed with the rpstblastn binary included in the blast+ package from the NCBI. Rpstblastn is run with an e-value of 1e-50. Outputs from Genemark.hmm and rpstblastn are parsed by Python scripts to generate gene predictions for the input DNA.

After gene prediction, protein sequences are individually queried against the nr database using blastp and an e-value of 1e-50 and up to 50 homologs are then retrieved. MUSCLE is then called with all default options. In order to retain the input order of the sequences, the stable.py script supplied with MUSCLE is then used to reorder the alignment. Because the standalone SIFT binary does not accept gaps in the aligned fasta input, any alignment columns with a gap in the reference are removed. Processed alignments are then analyzed using the info_on_seqs SIFT binary (SIFT version 5.0.3).

Results

Here we present EDSSI, a sequence analysis tool that inspects input DNA for potential syntax errors in the protein coding sequences when expressed in a bacterial context. By combining gene prediction, homolog retrieval, protein sequence alignment, and mutational analysis software, EDSSI predicts one type of genetic design error. EDSSI is available at <http://andersonlab.qb3.berkeley.edu/Software/EDSSI>.

Performance

The EDSSI analysis pipeline fits well into common design workflows. The two BLASTs are the main bottlenecks of the pipeline, so results for both are cached to improve performance for commonly queried sequences. The protein blast searches are also done in parallel to speed up performance. We timed the analyses for 20 *E. coli* genomic loci, and found that on average each kilobase takes 3 minutes to run.

Sequence inspector reports

EDSSI generates a report that contains a graphical representation of the input sequence. Detected ORFs are drawn as arrows and are labeled by a text annotation. The

ORF labels can be toggled on or off by a button at the top right of the display. ORFs are color-coded by level of evidence: ORFs with exact database matches are shown in green; ORFs with database hits but no exact match are shown in orange; and ORFs with no database hits are shown in black. Errors in the protein sequence are drawn as red vertical bars.

Each ORF links to its corresponding multiple sequence alignment. The sequence of the predicted ORF, or the input sequence, is given as the first sequence in the alignment. Subsequent protein sequences are ordered by similarity. The entire alignment is generated by JavaScript and also can be dynamically resized.

Predicted errors in the input sequence are depicted as color-coded vertical bars in the multiple sequence alignment, with more likely errors coded with a deeper shade of red. The amino acid characters are also color coded to facilitate visual comparison.

Several files from each analysis are available for download. The complete output data is available as a json download to facilitate interoperability with automated genetic design software. The aligned protein sequences are available in a fasta file. The conserved domain and homolog searches produce links to relevant, indexed abstracts in PubMed that users can read. The analysis pipeline also produces an annotated genbank file that can be read by popular DNA editors such as ApE, LaserGene, etc.

A literature search was conducted to find and evaluate underlying software for EDSSI. Genemark.hmm was found to correctly predict 93.5% of experimentally verified genes across a wide range of bacteria [21]. SIFT has been benchmarked against a large set of functional data from near complete mutagenesis of lacI, HIV protease, and T4 lysozyme. SIFT has false positive rates of 20% and false negative rates of 31% [18]. Finally, to test our indel addition to the SIFT analysis workflow, we selected ORFs from the EcoGene Verified Protein Starts set with an internal methionine, generated truncated protein sequences, and ran them through EDSSI. Truncated sequences may result in a false negative analysis if there are sufficient shorter, erroneous protein sequence entries in NCBI nr. However, EDSSI is able to predict 97% of the truncated protein sequences.

Three illustrative examples

To demonstrate the utility of EDSSI, we submitted the three published examples of syntax errors discussed above.

We first syntax-checked the synthetic construct pCTXvp60 [14]. The leucine-rich ORFs in pCTXvp60, including the toxicity causing ORF238, were correctly identified during the analysis. These artificial ORFs have no homologs and were therefore not identified by the conserved domain search. As expected, the artificial fusion protein CTXvp60, which lacks a bacterial RBS, was detected as two separate ORFs. However, knowing about the spurious ORF238 would likely have aided in troubleshooting the unexpected but observed plasmid toxicity in *E. coli*.

We next examined the X13303.1 *nif* cluster from [4] and EDSSI predicted 13 errors. The *nifS* gene is shown in Figure 3a as an example of how these point mutations are displayed. During *nif* cluster resequencing 18 non-synonymous mutations were found, of which 8 agree with the ones found by EDSSI. In comparison, the two homologous *nif* clusters found in had only 6 predicted errors in the 22kb region tested. Knowing about the predicted errors in the *nif* cluster sequence would likely have aided in

the debugging of the initial failed experiment. EDSSI also successfully identified the truncation in the published *invF* construct (Figure 3b). While some database entries share the same truncated translation start site, a majority of entries have the genuine start site. Knowledge of those entries would have been useful in the design of the experimental execution.

Discussion

Failed experiments are common in genetic engineering, and there is a need for software tools that provide suggestions for debugging these experiments. One common source of error is the subtleties in the DNA syntax of the tested constructs. Current practice relies on an implicit assumption that annotations and sequence databases are correct. However, as we have found in the published examples, those annotations can often mislead the engineer's thinking and can cause simple experiments to fail. By creating better tools for syntax checking and semantic verification, such experiments will have a lower chance of failure.

We used modern-day computer code editors as inspiration when we created EDSSI. Modern-day computer code editors can find syntax errors or warnings before runtime, enabling a faster debugging cycle. Similarly, our sequence inspector will allow for upfront handling of errors or can provide hypotheses for failed experiments.

Even though there does not exist a formal theory for how each side chain position contributes to overall protein function, statistical approaches for predicting deleterious mutations still can provide a means of prediction. By using the statistical techniques pioneered by programs like SIFT, our EDSSI output correlates with expert human analysis for the three published examples and our synthetic test sequences. However, in the *nif* gene cluster example, the sequence inspector did not identify all of the non-synonymous mutations found by resequencing. This disparity is either due to a false negative of the SIFT program, or some of the mutated positions are tolerated. Empirical testing on sequences substituted with each mutation for the desired nitrogen fixation function could differentiate between the two interpretations. The synthetic benchmarks suggest that the ORF prediction and gap scoring algorithms can be used for pre-experimental error prediction, while the amino acid substitution scoring may be useful for hypothesis generation in post-experiment debugging.

EDSSI can be improved by inclusion of more data. Analyses of rapidly evolving genes, such as endonucleases, or unique gene sequences will return many errors because the sequence inspector performs poorly when given few data points. As more strains and individuals are sequenced, the number of homologs for any given protein can be expected to increase. Also, more proteomic data will enable precise prediction of translation start sites. Integrating protein structure, when available, into the analysis could also improve predictions of effects of amino acid substitutions such as in the PolyPhen prediction pipeline [19]

For the biological engineer, the ability to rule out certain designs before fabrication will have an important role in enabling complicated designs [1, 20]. Already, with the right information, software systems can check the validity of designed logic gates [22] and metabolic pathways [23]. By addressing one common aspect of failure in genetic engineering, this tool will help move the practice closer to rational design.

Funding

The Synthetic Biology Engineering Research Center (SynBERC). National Science Foundation Graduate Research and Department of Defense National Defense Science and Engineering Graduate Fellowships (T.H.H.).

Conflict of interest statement: None declared.

Acknowledgements

The authors would like to thank Saurabh Srivastava, Oscar Westesson, Josh Kittleson, David Chen, and Zachary Russ for helpful discussions.

References

1. Kittleson, J.T., Wu, G.C., and Anderson, J.C. (2012). Successes and failures in modular genetic engineering. *Current Opinion in Chemical Biology*, **16**(3-4), 329-36.
2. Bork, P., and Bairoch, A. (1996). Go hunting in sequence databases but watch out for the traps. *Trends in genetics: TIG*, **12.10**, 425.
3. Farrer, R.A., Kemen, E., Jones, J.D., and Studholme, D.J. (2008). De novo assembly of the *Pseudomonas syringae* pv. *syringae* B728a genome using Illumina/Solexa short sequence reads. *FEMS microbiology letters*, **291.1**, 103-111.
4. Temme, K., Zhao, D., and Voigt, C.A. (2012). Refactoring the nitrogen fixation gene cluster from *Klebsiella oxytoca*. *Proceedings of the National Academy of Sciences*, **109.18**, 7085-7090.
5. Temme, K., personal communication.
6. Pati, A., Ivanova, N.N., Mikhailova, N., Ovchinnikova, G., Hooper, S.D., Lykidis, A. and Kyrpides, N.C. (2010). GenePRIMP: a gene prediction improvement pipeline for prokaryotic genomes. *Nat Methods*, **7**, 455-457.
7. Poptsova, M.S. and Gogarten, J.P. (2010). Using comparative genome analysis to identify problems in annotated microbial genomes. *Microbiology*, **156**, 1909-1917.
8. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. and Lander, E.S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241-254.
9. Wall, M.E., Raghavan, S., Cohn, J.D. and Dunbar, J. (2011) Genome Majority Vote Improves Gene Predictions. *PLoS computational biology* **7**, e1002284.
10. Stanke, M., Steinkamp, R., Waack, S. and Morgenstern B. (2004). AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic acids research* **32**, W309-W312.

11. Moon,T.S., Lou,C., Tamsir,A., Stanton,B.C. and Voigt C.A. (2012). Genetic programs constructed from layered logic gates in single cells. *Nature* **491**, 249-253.
12. Lillo,F. and Krakauer,D.C. (2007). A statistical analysis of the three-fold evolution of genomic compression through frame overlaps in prokaryotes. *Biol Direct* **2**, 22.
13. Lee,H., Whitaker,W. and Dueber,J.E. (2013). Avoidance of Truncated Proteins from Unintended Ribosome Binding Sites within Heterologous Protein Coding Sequences, in preparation.
14. Umenhoffer,K., Fehér,T., Balikó,G., Ayaydin,F., Pósfai,J., Blattner,F.R. and Pósfai,G. (2010). Reduced evolvability of Escherichia coli MDS42, an IS-less cellular chassis for molecular and synthetic biology applications. *Microbial cell factories* **9**, 38.
15. Powell,B.C. and Hutchison,C.A. (2006). Similarity-based gene detection: using COGs to find evolutionarily-conserved ORFs. *BMC bioinformatics* **7**, 31.
16. Edgar,R.C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* **32**, 1792-1797.
17. Ng,P.C. and Henikoff,S. (2001). Predicting deleterious amino acid substitutions. *Genome research* **11**, 863-874.
18. Ng,P.C., and Henikoff,S. (2006). Predicting the effects of amino acid substitutions on protein function. *Annu. Rev. Genomics Hum. Genet.* **7**, 61-80.
19. Adzhubei,I.A., Schmidt,S., Peshkin,L., Ramensky,V.E., Gerasimova,A., Bork,P., Kondrashov,A.S. and Sunyaev,S.R. (2010). A method and server for predicting damaging missense mutations. *Nature methods* **7**, 248-249.
20. Purnick,P.E. and Weiss,R. (2009). The second wave of synthetic biology: from modules to systems. *Nature Reviews Molecular Cell Biology* **10**, 410-422.
21. Hyatt,D., Chen,G.L., Locascio,P.F., Land,M.L., Larimer,F.W. and Hauser,L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics* **11**, 119.
22. Bilitchenko,L., Liu,A., Cheung,S., Weeding,E., Xia,B., Leguia,M., Anderson,J.C. and Densmore,D. (2011). Eugene—a domain specific language for specifying and constraining synthetic biological parts, devices, and systems. *PloS one* **6**, e18882.
23. Hatzimanikatis,V., Li,C., Ionita,J.A., Henry,C.S., Jankowski,M.D. and Broadbelt,L.J. (2005). Exploring the diversity of complex metabolic networks. *Bioinformatics* **21**, 1603-1609.

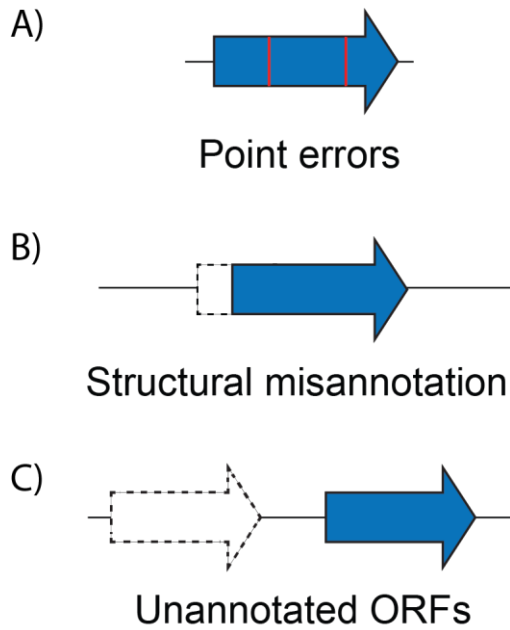


Figure 1. Causes of syntactic error in genetic designs. (A) Point errors can result from erroneous sequencing or manual data entry. (B) Structural misannotation caused by late start sites can result in N-terminal truncations. (C) Unannotated ORFs can result in the expression of unintended genes.

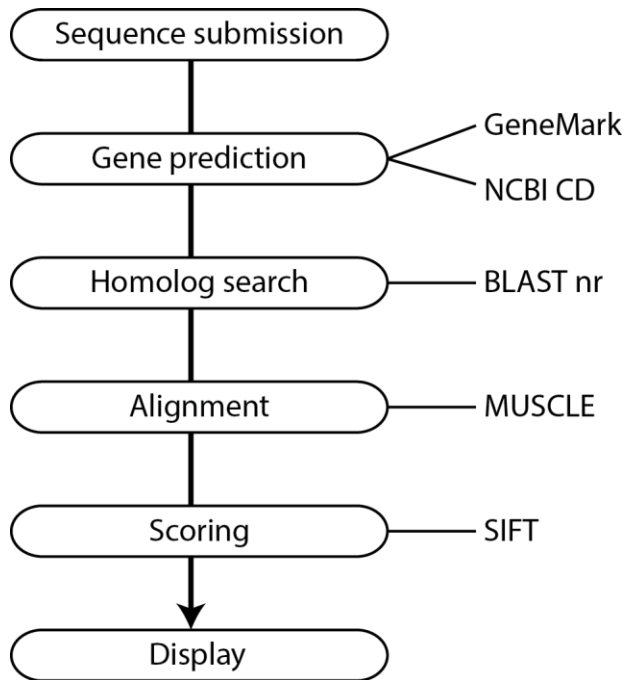


Figure 2. Data analysis workflow. The user submits DNA sequences through an online interface, which are then run through gene prediction, homolog search, alignment, and scoring. Associated programs or algorithms for each stage are shown.

A)

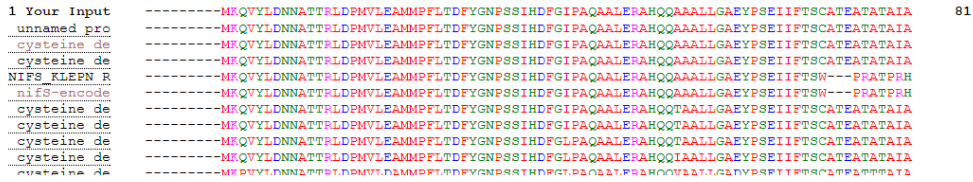
Click ORFs to see their alignment



B)

Error Bar Legend: Error likely Unlikely

ORF Legend: Exact Match found Matches found, but no exact hit No match found



Click ORFs to see their alignment



Error Bar Legend: Error likely Unlikely

ORF Legend: Exact Match found Matches found, but no exact hit No match found

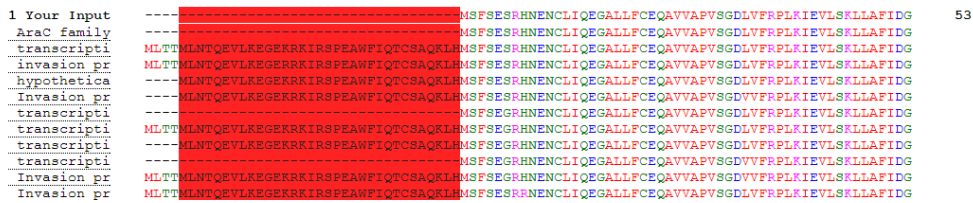


Figure 3. Examples of the results page. The top panel depicts point mutations in the genbank entry X13303 used in the study [4]. The bottom panel depicts the truncated invF gene used in the study [11].

Multiplex Ortholog Library Synthesis and Expression Testing

Abstract

Real-world synthetic biology is still very challenging, with projects taking years to complete. Building genes and testing candidates are both expensive. To that end, we have developed Multiplex Ortholog Library Synthesis and Expression Testing (MOLSET) for rapid building and expression characterization of many genes as a pool. In multiplex ortholog library synthesis, we have developed a scalable and inexpensive method for directly synthesizing an entire library of genes from microarray oligonucleotide pools. We then used a genetic reporter for multiplex expression characterization and demonstrate the feasibility of our approach by building and testing 90 genes for empirical evidence of soluble expression.

Used concurrently, these techniques allows an engineer to rapidly build and screen the expression of a library of genes to limit the number of devices going into an assay. Both techniques are highly scalable; thus their accessibility and cost will track with the improvements in the underlying technologies of multiplex DNA reading and writing.

Preface

The contents of this chapter are based on an ongoing project in the laboratory of Dr. J. Christopher Anderson. My contributions to this work included developing the MOLSET methodology, writing in-house scripts for nucleotide sequence design, and writing this chapter. Dr. Phillip Elms, Dr. David Sukovich, Tahoura Samad, Tobias Stritmatter, Dr. J. Christopher Anderson, Saurabh Srivastava, Paul Ruan, and Bo Curry contributed to the work in this chapter or general project coordination and management.

Introduction

The design, build, test cycle paradigm has emerged as a powerful framework for the engineering of biology. Recently, we have developed a design tool called the Act Synthesizer, a system for predicting biosynthetic pathways for desired products. The Act Synthesizer uses a curated database of observations to provide the engineer with a list of enzymatic transformations and the associated gene sequences. However, the build and test stages for constructing and assaying genetic parts in the laboratory are still limiting. Currently, building genes requires access to source organisms or the use of expensive gene synthesis. Even after genetic material is in hand, assays are often laborious or expensive, resulting in a need for prioritization of which genes to test. One such metric for prioritization is soluble expression. By first prioritizing genetic sequences that are solubly expressed, engineers can potentially avoid assaying recombinant proteins that end up in non-functional aggregation bodies [1-4]. However, currently soluble expression has been used as a gene ranking metric as most assays are laborious or require access to specialized equipment.

To address these limitations we report herein Multiplex Ortholog Library Synthesis and Expression Testing (MOLSET), a high-throughput gene synthesis and expression assay method. In a multiplex manner, MOLSET synthesizes up to a hundred genes from microarray oligos and assays their soluble expression using a recently-developed indirect folding reporter in conjunction with next-generation sequencing. The method uses only common one-pot techniques, can be implemented in any molecular biology laboratory, and is able to synthesize

the majority (>95%) of designed genes. The results of MOLSET can be used to prioritize testing of genes, and we show how incorporation of that data impacts the predictions from the Act synthesizer. Together, these technologies are highly scalable and their accessibility and cost will only advance with improvements to the underlying technologies.

Results

To improve the ease of building genetic sequences, we set out to develop an inexpensive and easy-to-use gene synthesis method. Previous efforts have shown that gene synthesis from microarray oligos is possible [5, 6]. However, previous approaches have been difficult to scale and not widely accessible for the following reasons: 1) they use complicated protocols, 2) require the use of robotics to be practical, and 3) have errors that require clonal enzymatic correction or sequencing. Points 1) and 2) result from demultiplexing the material at the oligo or gene stage and require the handling of many samples in parallel. Our approach to simplify the protocol by demultiplexing the complex pool as late as possible, if at all. Thus, we set out to make a robust protocol to make genes at the hundred-scale from OLS pools with only one-pot molecular biology techniques.

A robust method for Ortholog Library Synthesis

To enable multiplex ortholog library synthesis, we first standardized the length of synthesized sequences and computationally minimized sequence crosstalk. For each design, synthon lengths were standardized by addition of random padding sequence (GC 50%). Codons were randomly chosen with a choice weight proportional to *E. coli* codon frequencies and oligos were designed to have 15 nt of unique sequence at both termini. Synthons for each gene were designed to be 711 bp. Gene-specific primer sequences and universal primer sequences were then added on the end of each sequence [6, 7, 16]. The sequence and oligo design constraints are illustrated in Figure 1.

To empirically test if we could directly synthesize genes from OLS pools, we designed a chip for 83 GFP family members for a total of 2,700 oligos and 60.6 kb of designed sequence. Oligos were synthesized by Agilent and received as a multiplex pool. After phosphorylation and a one-pot ligation, we performed gene-specific PCRs with orthogonal primer pairs and subcloned products into an expression vector. The robustness of the protocol was demonstrated when correctly-sized PCR product was observed even when 1 picomole of a 25 nt random oligo (N_{25}) was doped into the ligation reaction.

Correct gene synthesis products for 80 genes were found by screening cloned products by fluorescence or clonal sequencing. The frequency of correct colonies determined by sequencing naive clones was determined to be 15% (6/40), with the majority of errors being point deletions, as is expected from the oligo synthesis methodology [8]. We next tested designs with more genes or longer synthons; however, we were unable to assemble 1000 genes of 800 bp (unable to get any perfect sequences for a majority of genes) or 200 genes of 1.5 kb (partial fragments only). Thus we continued with the proven design of a hundred synthons of 800-900 bp in length.

Multiplex Ortholog Synthesis

We next sought to enable a one-pot process by eliminating the need for gene-specific amplification. Our initial attempts to directly amplify the pool of assembled genes from the complex ligation reaction using conventional PCR yielded only short (100-200 bp) products. We reasoned that the shorter products were favored by the PCR reaction and sought to counter or invert the length bias of PCR. We tried amplification by emulsion PCR [9] and suppression PCR [10]. Water-in-oil emulsion PCR was carried out on the ligation reaction and was able to generate a faint band of the correct size. Use of emulsion PCR alone was found to be not robust as sometimes no band was seen after amplification. We then used emulsion PCR to add on inverted repeats on the end of amplicons that act as suppression tails. A single primer which binds to the inverted repeats at the end of the amplicon is used in suppression PCR for amplification. The inverted repeats can anneal to each other and compete with primer binding [10]. Shorter amplicons exhibit the suppression effect more than longer amplicons, thus suppression PCR is biased towards longer amplicons. Suppression PCR product was then cloned directly into an expression vector and colonies were randomly picked and sequenced. The number of correct, full-length genes was 21% (3/14); however, 43% (6/14) of clones were fusions of two genes and the remainder of the errors had deletions or were truncated genes. We concluded that Multiplex Ortholog Synthesis was sufficient to quickly create a multiplex pool of genes with an acceptable error rate, and next sought an appropriate downstream assay for expression.

A high-throughput expression assay based on Tat quality control

We sought to develop a multiplexable expression assays in order to avoid individual cloning and sequence verification of genes. Recently, a selection assay based on the twin-arginine export quality control mechanism has been developed [11]. In this system, the gene of interest (GOI) is fused at the 5' end with a Tat export signal derived from trimethylamine N-oxide reductase (ssTorA) and fused at the 3' end with the mature TEM-1 beta-lactamase. Previous work has shown that translocation of the fusion protein and conferral of a resistance to ampicillin depends on the correct folding of the gene of interest [12]. We adapted this system in conjunction with Next Generation Sequencing to develop our expression assay, as shown in Figure 2.

We then designed a Multiplex Ortholog Library Synthesis pool for 95 genes, with 7 genes being negative controls previously shown to be poorly expressed in *E. coli*, 6 *E. coli* genes as positive controls, 1 engineered monomeric GFP also as a positive control, and 69 genes chosen from our Act system. These genes were randomly chosen in order to have a wide range of assay in order to assess the performance of the multiplex expression assay. These genes were synthesized and cloned into the pSALect-EB vector. The library was electroporated and a diversity of 5×10^5 was observed by titer plating.

We plated cells on 1, 2.5, 5, or 10 $\mu\text{g/mL}$ of ampicillin and chloramphenicol and found that 1 $\mu\text{g/mL}$ of ampicillin yielded no drop in titer compared to chloramphenicol only. In contrast, we observed a 10% survival in titer on 2.5 or 5 $\mu\text{g/mL}$ of ampicillin, which was on par with the expected survival rate based on Sanger sequencing of naive clones. By sequencing clones grown from the 5 $\mu\text{g/mL}$ ampicillin plate, we observed that 60% (27/45) were full length and correct, suggesting that the Multiplex Expression Assay system can also be used for multiplex, non-enzymatic gene synthesis error correction.

We next sought to characterize in a multiplex manner the expression of the synthesized genes and plated approximately 10^8 cells on solid media supplemented with 5, 10, 50, or 100 $\mu\text{g}/\text{mL}$ of ampicillin and 25 $\mu\text{g}/\text{mL}$ of chloramphenicol. Plates were incubated at 30°C overnight and we observed a titer of 10% for 5 $\mu\text{g}/\text{mL}$, 3% for 10 $\mu\text{g}/\text{mL}$, 1% for 50 $\mu\text{g}/\text{mL}$, and 0.3% for 100 $\mu\text{g}/\text{mL}$. Then, we scraped the plates, minipreped to recover the plasmid library, prepared sequencing samples using a TruSeq kit, and performed sequencing with a MiSeq instrument. A total of 7.9 million reads were generated and were mapped to the reference genes. Overall, 18.7% of reads were successfully mapped to a designed gene, with the minimal per-pool mapping rate being 17.6% and the maximal mapping rate 20.2%.

The TruSeq random fragmentation method results in shotgun coverage of the plasmids and thus the full insert is not sequenced in any single read. Amplicon-based approaches would give rise to skewed counts as inserts are of different sizes in our assay. Accordingly, partial gene fragments show up in our mapping. However, from differences in the base-by-base coverage we can detect if coverage count arises from subfragments. By taking the median of the base-by-base coverage, we have a length-normalized count of representation. For the 5 $\mu\text{g}/\text{mL}$ condition, the median read count correlated with Sanger sequencing data. The same gene, Brenda92, was found to be the most represented in both the Sanger and the Illumina sequencing. Ratios between the most represented and the second ranked genes were comparable (8:3 clones with Sanger sequencing versus 3197:1131 using Illumina median read coverage). We then normalized counts by the pool to arrive at a dimensionless number that can be compared across different ampicillin conditions.

Our data showed that we synthesized 96% of the genes, and also enabled inferring gene expression. Expression was then computed by taking the pool-normalized values for each gene and normalizing to the 5 $\mu\text{g}/\text{mL}$ condition. We found that all of the negative solubility controls exhibited a pattern of having low representation in the higher ampicillin concentrations, rapidly falling off after 5 $\mu\text{g}/\text{mL}$. Of the GFP family members, we included a well-folding, monomeric positive control, mKG [13], and found that it was one of the most represented genes in the 100 $\mu\text{g}/\text{mL}$ pool. Of the 6 *E. coli* controls, we found that two survived at high ampicillin concentrations, while the other 4 died at 50 $\mu\text{g}/\text{mL}$ or more of ampicillin. These results suggest that while multiplex expression assaying using the tripartite fusion system is convenient, it suffers from false negatives.

Correlation with confirmatory experiments

In order to confirm our expression sequencing analysis, we dilution spotted overnight cultures of 6 retransformed clones onto plates of 0-400 $\mu\text{g}/\text{mL}$ of ampicillin to confirm the phenotypes. The dilution plating agreed with our NGS findings and can be seen in Figure 3. To independently confirm expression of the 6 representative samples, we expressed them as FLAG-tagged chimeras and performed Western blotting to look at soluble versus insoluble fractions. The Western blot generally correlated with ampicillin growth (data not shown), but there was one exception, a dimer which expressed solubly as a FLAG-tag fusion, again suggesting that the multiplex expression system suffers from false negatives.

NGS-predicted expression was then correlated with text mined expression predictions in the Act Ontology. Evidence for expression was inferred from the "Cloned" commentary section of the BRENDA database. We then searched for comments with the organism name "escherichia", and without the terms "inclusion bodies" and "folding", as evidence of expression.

Of the 69 BRENDA-derived test genes, 58 (84%) were predicted by text mining to be expressed in *E. coli*. NGS predictions using a 10% representation cutoff (gene pool representation should be more than 10% of the pool representation observed in the 5 µg/mL condition) predicted 43 (62%) genes to be expressable.

Methods

Design of synthesized sequences

Sequences of all GFP family members in Uniprot were downloaded and a phylogenetic tree was made. We then selected GFP family members to synthesize from this tree. Protein sequences were converted to nucleotides using a weighted random codon algorithm design in-house. Oligonucleotides were subsequently designed from the nucleotide sequences with the following constraints: no longer than 175 nt and no more than a 15 nt exact match between two oligos at either terminus.

Gene synthesis by high-temperature ligation

OLS oligos were synthesized by Agilent and received resuspended in 100 µL of TE buffer. Oligos were phosphorylated using 3 µL T4 Ligase Buffer (NEB), 24 µL OLS oligos, and 3 µL T4 PNK(NEB) for 37°C at 1 hour. The reaction was heat inactivated at 65°C for 30 min and held at a final 16°C. Testing several commercially available thermostable ligases revealed no differences in the gene-specific PCR for a subset of 20 genes. We used 9 degrees North ligase (NEB) for all experiments.

Whole pool ligation was performed with 12 µL phosphorylated oligos, 4 µL 50% 3350 PEG (Carbowax P146-3), 2 µL 9°N™ buffer, 2 µL (80 U) of 9°N™ ligase (NEB). Reactions were performed in a MJ Research PTC-200 thermocycler using the following program: 95°C for 2 minutes, 65°C for 24 hours, and 4°C hold. The ligation product was used as template for gene-specific PCRs or emulsion PCR. Gene-specific PCR was performed using 0.25 µL of the ligation product as template with gene-specific primers. Emulsion PCR is detailed in the following section.

Emulsion PCR for post-ligation amplification

Emulsion oil mix was prepared with 450 µL Span 80 (Fluka 85548), 40 µL Tween 80 (Sigma P4780), 5 µL Triton X-100 (Promega H5142), and 9505 µL mineral oil (Sigma M5904) as described in [9]. Oil was thoroughly vortexed to mix. Separately, PCR reaction mix was prepared on ice using Q5 polymerase (NEB). PCR reactions were performed using 10 µL of ligation product as template supplemented with 0.5 µL (1 U) of Q5 polymerase, 20 µL Q5 reaction buffer, 1 mM dNTP, and water for a total reaction volume of 100 µL.

For emulsification, PCR reaction was mixed with oil at a 1:10 (PCR:oil) volumetric ratio. PCR mix was pipetted into a cryovial tube containing emulsion oil and vortexed at maximum power using a VWR benchtop vortexer for 1 minute until a milky white emulsion formed. The emulsion was distributed as 100 µL aliquots and PCR was performed in a MJ Research PTC-200 thermocycler.

To break the emulsion post-PCR, aliquots were consolidated into microcentrifuge tubes and spun for 20 minutes to separate the oil and aqueous phases. As much oil as possible was

pipetted off the top of the biphasic solution, 300 μ L of 2-butanol (Sigma-Aldrich 19440) was added to break the emulsion, and tubes were vortexed. For reaction clean-up, 1 mL of ADB (Zymo) was added, tubes were vortexed, and PCR clean-up columns (Zymo research) were used to purify the amplicons.

Suppression PCR to generate clonable amplicons

Purified emulsion PCR product was used as a template for suppression PCR. Both emulsion primers were designed with a suppression tail of (CATCAGGTTTCATCCTGCCGGCATGAGCGGCTAACGG) so that amplicon ends form an inverted repeat. For suppression PCR, the distal-binding primer (CATCAGGTTTCATCCTGCCGG) was used (30 cycles, T_m of 55°C). PCR products were visualized on a gel and the band of the appropriate length was excised and cloned into a multiple cloning site flanked by EcoRI and BamHI.

Solubility assay using a beta-lactamase folding reporter

Matthew DeLisa graciously provided the pSALect vector. We modified the pSALect vector to create pSALect-EB by placing EcoRI and BamHI restriction sites in between the tat signal sequence and the mature TEM-1 beta-lactamase sequence. For library creation, digested pSALect-EB vector and amplicons were ligated and purified with a PCR clean-up column (Zymo). Purified DNA was then electroporated into MC1061 derivative strains [14]. Cells were rescued for 2 hours at 37°C and grown overnight in 200 mL 2YT liquid media supplemented with 25 μ g/mL chloramphenicol. Rescued cells were also dilution plated onto LB chloramphenicol plates for titering.

A dilution equivalent of 1 μ L overnight culture was then plated on LB plates supplemented with chloramphenicol (25 μ g/mL) and ampicillin at different concentrations ranging from 1 to 100 μ g/mL. Plates were incubated for 16 hours at 30°C. Plasmids were harvested from plates as libraries or colonies were grown overnight for minipreps.

DNaseq of Libraries

Library minipreps were quantitated using a Nanodrop (Thermo Scientific) and fragmented using a Covaris S220 using the recommended protocols in the TruSeq kit (Illumina). The TruSeq procedure was used to prep the libraries for sequencing. Pools were prepared separately, barcoded, quantitated using a Library Quant Kit (Kapa Biosystems), combined, and sequenced on a MiSeq using a 300 cycle v2 kit.

Reads were quality trimmed and mapped to the reference sequences using BWA (0.6.1-r104). Samtools mpileup was used to extract per-base coverage and then an in-house python script and Microsoft Excel were used to normalize the data. As some inserts were partial gene fragments and also contribute to the per-base read coverage score, we took the median as the read coverage score for the entire gene. Manual inspection of the read coverage for several genes showed that the median was an acceptable measurement of whole gene read coverage. The read coverage per gene is then pool-normalized by dividing by the sum of read coverages for all genes in each pool.

Confirmation of solubility with Western blotting

Library minipreps from the 5 µg/mL ampicillin condition were digested and the 700 bp band corresponding to library of genic inserts was gel purified and cloned into an arabinose-inducible expression vector with a C-terminal 3x FLAG tag. 72 colonies were Sanger sequenced and 26 unique inserts were recovered. Minipreps were retransformed and strains were grown overnight, reinoculated, induced with arabinose (0.2% w/v), and harvested after 4 hours. Cells were pelleted by centrifugation for 5 minutes at 2500 rcf, and the cell pellet was resuspended with BugBuster MasterMix (Novagen) at a ratio of 1 mL BugBuster per 0.1 g cell pellet. Cells were lysed for 20 minutes at 25°C on a rocking platform and soluble protein was recovered by taking the supernatant after centrifugation at 12,000 rcf for 15 minutes. The insoluble fraction was resuspended in an equal amount of BugBuster. Subsequent Western blotting was performed with Monoclonal Anti-FLAG M2-HRP antibody (Sigma A8592) and ECL Western Blotting Substrate (Pierce 32106). Images were quantitated using ImageJ.

Discussion

We have developed new methodologies, called MOLSET, for rapidly building genes and prioritizing them for downstream assays based on soluble expression. For building genes, we simplified the process of creating genes from complex microarray oligo pools. Our method requires only one-pot reactions and is able to create genes with length 811 bp at the hundred-scale with a high rate of success (>95% of genes). These results enable researchers to perform only 5 unit operations to make a pool of 100 genes as opposed to hundreds of unit operations required by previous protocols. Additionally, with the use of the beta lactamase fusion system and selecting on ampicillin, we have shown it is possible to perform non-enzymatic gene synthesis error correction in a pooled format. Taking these steps in combination, even if the eventual goal is to generate clonal inserts, our methods could save researchers much labor.

Our motivation in developing the expression testing in MOLSET is based on two observations. Firstly, expression issues can negatively impact recombinant protein function. Secondly, downstream functional assays can be limited in throughput. In combination, these two observations point towards a need for cataloguing of the “expressability” of selected gene sequences such that 1) poorly expressed genes can be avoided in the design stage or 2) genes predicted to express can be prioritized in assays. Our multiplex expression test uses the Tat-pathway export of a beta-lactamase folding reporter in conjunction with next-generation sequencing to quickly assay expressability of hundreds of genes using the widely available methods of plate-based selections. Previously, tripartite fusions in conjunction with NGS have been used to annotate genomes [15] and other reports have used NGS for characterization of degron mutants [17]. Our study extends on those results and show that it is possible to use NGS to read out multiplex expression measurements.

Using the Tat export pathway allows our expression system to avoid false positives arising from translation due to spurious ribosomal binding sites internal to the assayed ORF, however, it also confers some disadvantages on our system. The Tat pathway has a limit on the size of proteins it can export, and while proteins of up to 120 kDa have been shown to be exported [12], all of the factors that influence what can and cannot be exported are

unknown. Taken together, these and our findings suggest that the Tat β -lactamase system has few false positives, and a false negative rate that can be complemented by other approaches.

Complementary approaches would either use a different method of fusing GOIs to the β -lactamase reporter or the use of other reporters. There has been a loop insertion system developed for beta lactamase [22]. Also, the use of GFP as a expression reporter has been shown to correlate with the solubility of the fusion partner [18]. Currently, there are also split or circular permutation GFP folding assays which only require a tag of as little as 15 amino acids [19-21]. In conjunction with fluorescence-activated cell sorting and DNA sequencing, it would be possible to do multiplexed expression assays with a broad range of organisms and genes.

With affordable gene synthesis now available, there can be a mismatch between the throughput at the building stage and what can be assayed. In the absence of a genetic selection, downstream assays such as mass spectrometry are limited by their accessibility, throughput and cost. One solution to this mismatch is to pre-screen genetic designs for easily assayable features that can influence final performance. We have developed a multiplexable assay for expression in this report, but future development of multiplexable assays for aspects such as protein-protein interactions or impact on cellular fitness will also enable a more predictable approach to genetic engineering.

References

1. Zhou, Kang, et al. "Enhancing solubility of deoxyxylulose phosphate pathway enzymes for microbial isoprenoid production." *Microbial cell factories* 11.1 (2012): 148.
2. Steen, Eric J., et al. "Metabolic engineering of *Saccharomyces cerevisiae* for the production of n-butanol." *Microb Cell Fact* 7.1 (2008): 36.
3. Atsumi, Shota, et al. "Engineering the isobutanol biosynthetic pathway in *Escherichia coli* by comparison of three aldehyde reductase/alcohol dehydrogenase genes." *Applied microbiology and biotechnology* 85.3 (2010): 651-657.
4. Yoshikuni, Yasuo, et al. "Redesigning enzymes based on adaptive evolution for optimal function in synthetic metabolic pathways." *Chemistry & biology* 15.6 (2008): 607-618.
5. Borovkov, Alex Y., et al. "High-quality gene assembly directly from unpurified mixtures of microarray-synthesized oligonucleotides." *Nucleic acids research* 38.19 (2010): e180-e180.
6. Kosuri, Sriram, et al. "Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips." *Nature biotechnology* 28.12 (2010): 1295-1299.
7. Xu, Qikai, et al. "Design of 240,000 orthogonal 25mer DNA barcode probes." *Proceedings of the National Academy of Sciences* 106.7 (2009): 2289-2294.
8. LeProust, Emily M., et al. "Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process." *Nucleic acids research* 38.8 (2010): 2522-2540.
9. Williams, Richard, et al. "Amplification of complex gene libraries by emulsion PCR." *Nature methods* 3.7 (2006): 545-550.
10. Shagin, Dmitry A., et al. "Regulation of average length of complex PCR product." *Nucleic acids research* 27.18 (1999): e23-i.
11. Fisher, Adam C., Woojin Kim, and Matthew P. Delisa. "Genetic selection for protein solubility enabled by the folding quality control feature of the twin-arginine translocation pathway." *Protein Science* 15.3 (2006): 449-458.
12. Lim, Hyung-Kwon, et al. "Mining mammalian genomes for folding competent proteins using Tat-dependent genetic selection in *Escherichia coli*." *Protein Science* 18.12 (2009): 2537-2549.
13. Ueyama, Takehiko, et al. "Sequential binding of cytosolic Phox complex to phagosomes through regulated adaptor proteins: evaluation using the novel monomeric Kusabira-Green System and live imaging of phagocytosis." *The Journal of Immunology* 181.1 (2008): 629-640.
14. Kittleson, Joshua T., Sherine Cheung, and J. Christopher Anderson. "Rapid optimization of gene dosage in *E. coli* using DIAL strains." *Journal of biological engineering* 5.10 (2011).
15. D'Angelo Sara, Velappan Nileena, et al. "Filtering" genic" open reading frames from genomic DNA samples for advanced annotation." *BMC Genomics* 12.
16. Zhou, Xiaochuan, et al. "Microfluidic PicoArray synthesis of oligodeoxynucleotides and simultaneous assembling of multiple DNA sequences." *Nucleic acids research* 32.18 (2004): 5409-5417.
17. Kim, Ikjin, et al. "High-throughput Analysis of in vivo Protein Stability." *Molecular & Cellular Proteomics* 12.11 (2013): 3370-3378.
18. Waldo, Geoffrey S., et al. "Rapid protein-folding assay using green fluorescent protein." *Nature biotechnology* 17.7 (1999): 691-695.
19. Cabantous, Stéphanie, Thomas C. Terwilliger, and Geoffrey S. Waldo. "Protein tagging

- and detection with engineered self-assembling fragments of green fluorescent protein." *Nature biotechnology* 23.1 (2004): 102-107.
20. Blakeley, Brett D., Alex M. Chapman, and Brian R. McNaughton. "Split-superpositive GFP reassembly is a fast, efficient, and robust method for detecting protein–protein interactions in vivo." *Molecular BioSystems* 8.8 (2012): 2036-2040.
 21. Cabantous, Stéphanie, et al. "New molecular reporters for rapid protein folding assays." *PLoS One* 3.6 (2008): e2387.
 22. Foit, Linda, et al. "Optimizing protein stability in vivo." *Molecular cell* 36.5 (2009): 861-871.

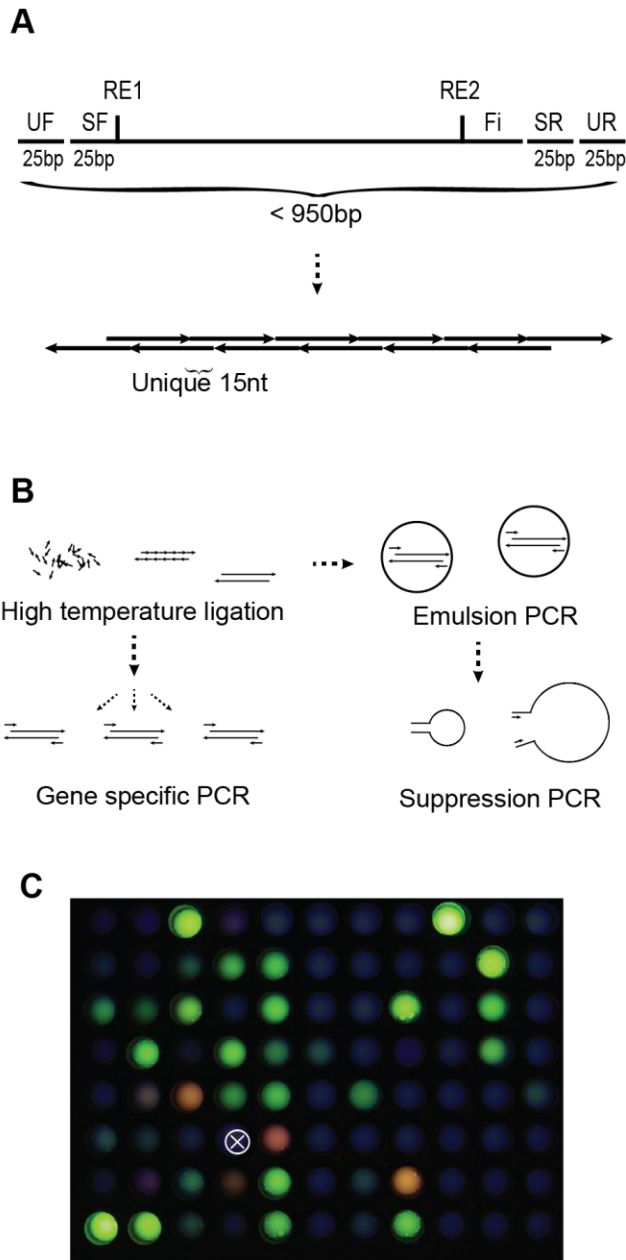


Figure 1. Multiplex Ortholog Library Synthesis. A) Nucleotide sequence and oligo design constraints. Synthons are not to exceed 950 bp including Universal Forward and Reverse (UF, UR, respectively) sequences, Specific Forward and Specific Reverse (SF, SR, respectively), Restriction Endonuclease (RE) recognition sites, and filler (Fi). Oligos are designed such that all terminal 15 nucleotides are unique. B) Overview of the gene synthesis process. Oligos are phosphorylated and subjected to a high temperature ligation. Next, individual genes can be amplified via specific PCR, or emulsion PCR and suppression PCR can be performed to amplify genes in a one-pot, multiplex manner. C) Cells expressing 88 GFPs synthesized in this study. 1 mL of cells were grown overnight, concentrated by centrifugation, and resuspended in PBS in a clear bottom 96-well plate with opaque siding. Images are combined from UV illumination (top

layer, 50% transparency) and blue light illumination. Well F4 contains empty media and is marked with a white X.

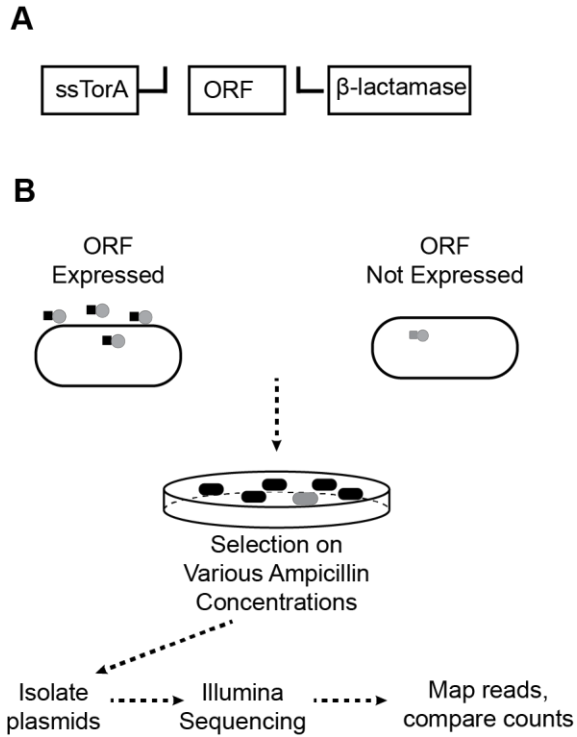


Figure 2. Overview of multiplex expression testing. A) Test ORFs are fused to a N-terminal Tat pathway secretion signal (ssTorA) and a C-terminal TEM-1 β -lactamase. B) Solubly expressed ORFs result in mature β -lactamase fusions exported to the periplasm of the cell. The library is then selected on various ampicillin concentrations and relative abundances are assayed with Illumina sequencing.

Chapter 3

The Act Synthesizer: Design and Testing of Biosynthetic Pathways

Abstract


The microbial production of chemicals is an active area of investigation that has not been characterized yet in terms of its opportunities, limitation, and risks. Contrary to the common perception that the avenues for biosynthesis of chemicals is unquantifiable, here we present as a conservative over-approximation an enumerated set of a few thousand chemicals that can be biologically produced, given current electronic enzymatic datasets. Not only that, for each of these “reachable” chemicals one can exhaustively characterize the set of heterologous pathways that can lead to it, if engineered into the cell. This enumerated set of biosynthesizable chemicals can also be screened for bioactivity, leading to an algorithmic and quantifiable characterization of biosafety concerns and potential mitigation strategies.

Preface

The contents of this chapter are based on an ongoing project. My contributions to this work included designing experiments, assisting in the development of the Act system, performing wetlab experiments, and writing portions of this chapter. Dr. Saurabh Srivastava is the main developer of the Act system and also wrote portions of this chapter. Paul Ruan, Dr. J. Christopher Anderson, Dr. Sanjit Seshia, and Dr. Ras Bodik contributed to the work in this chapter or general project coordination and management.

Introduction

Biosynthetic production of chemicals, using Microbial Chemical Factories (MCFs), is an important area of synthetic biology research. Exemplar instances exist of optimized production strains for precursors to drugs [1], polymers [2], and fuels [3]. To date, typical pathway design processes involve copying natural biosynthetic pathways or manual searching through literature to back-track possible biosyntheses step-by-step to known metabolites. However, these “manual” methods require specialized knowledge and can miss finding pathways as the number of possible biosyntheses to consider can become quite large after only a few steps.

Prior approaches have demonstrated that computational biosynthesis design can supplement or surpass manual design capabilities [2,4,5] and enables the biological production of novel chemicals.  [\[1\]](#) In this work, we develop a data-driven MCF-design system, called the Act Synthesizer, and exhaustively enumerate the current opportunities and limitations of the biosynthesizable chemical space. We additionally experimentally validate Act Synthesizer designed pathways by engineering *E. coli* to produce the household painkiller, acetaminophen.

Results

Approach: We use a graph-based approach to represent the known or reachable metabolic space. In our bipartite graph representation, there are two types of nodes: reactions and chemicals. Common “cofactors” such as H₂O, NADH, or ATP are manually defined and ignored in edge construction. To design novel biosynthetic pathways, we first define a core set of chemicals present in the host (see definition of L0, below) and then use a breadth-first search to enumerate enzymatic steps leading to new chemicals. Pathways found using our simple node-edge approach are next subjected to a multi-stage process that considers “ranking factors” such as 1) if

all substrates are present for multi-substrate/product reactions and 2) known directionality of the reaction. Act Synthesizer algorithm details are available in the Methods section.

To populate the Act system, we have used the BRENDA database [7], a repository of enzymes and their catalyzed biochemical reactions.

Reachables: We have categorized biosynthesizable chemicals, here called reachables, based on the difficulty of their production within a biological chassis. Level 0 (L0): endogenous chemicals natively produced in the organism, here taken as a list of 192 *E. coli* metabolites [6]. Level 1 (L1): chemicals naturally produced by another organism. Level 2 (L2): chemicals that need enzymes from multiple different organisms. Level 3 (L3): chemicals that require a speculated enzymatic substrate specificity. Level 4 (L4): chemicals that need an enzyme with engineered substrate specificity reaction, Level 5 (L5): chemicals that need a reaction step with unnatural chemistry. Levels L3-L5 require predictive models for enzyme function (e.g., predicting substrate binding via Structure Activity Relationships) and we leave that for future work. This report characterizes the L0-L2 MCF space.

The Act Synthesizer can fully enumerate the L0-L2 chemicals, given current databases of enzymatic catalysis. Using a list of metabolites native to *E. coli* as the L0 set, the L0-L2 space contains 3519 chemicals. The distribution of number of reachables by enzymatic steps is shown in Figure 1. By cross-referencing with other datasets such as the Sigma-Aldrich or Drugbank catalogs, we can highlight biosynthesizable polymers, therapeutic drugs, or chemical building blocks. We next enumerated all possible enzymatic pathways for each reachable chemical.

Enzymatic cascades: Given a reachable chemical, all possible enabling enzymatic routes can be expressed in the form of an acyclic directed graph. This acyclic directed graph starts with native host metabolites and ends at the reachable chemical. We call these set of possibilities the *enzymatic cascade* to that chemical. A simplified schematic of an enzymatic cascade is in Figure 2. [12] Square nodes indicate intermediates, and round nodes represent reactions. Within this cascade, any set of enzymes that make a contiguous connected sequence from natives to reachable is a viable MCF pathway to the target chemical. The cascade can be ranked by a confidence metric for success relative to other pathways through the cascade. The metric used for confidence evaluation is enzyme expression data within that host and the number of other enzymes doing similar catalysis (but which potentially work on substantially different substrates.) The set of L0-L2 reachables and their cascades provide microbial engineering options for a variety of commonly targeted MCF chemicals, including adipic acid, glucaric acid, squalene, amongst others. In addition, the reachables also contain some unnatural chemicals whose biosynthesis has not yet been demonstrated. One such reachable target is 4-acetaminophen (also called paracetamol or Tylenol[®]). To test the validity of Act's predictions of its reachability and the corresponding cascade we engineered an *E. coli* strain with the suggested pathway.

Confirmation of Act-designed pathways: We chose to validate Act designs by implementing the pathway shown in Figure 3 for 4-acetaminophen. This L2 pathway branched off from chorismate biosynthesis and was unnatural in two regards. First, it suggested a gene from the common mushroom (*Agaricus bisporus*) for conversion of PABA to 4-aminophenol for the penultimate step. Second, it suggested using 4-aminophenol as an unnatural substrate for the *E. coli* gene N-

hydroxyarylamine O-acetyltransferase (NhoA). The enzyme's natural substrate is p-aminobenzoic acid but it has also been shown to acetylates 4-aminophenol to produce acetaminophen [8]. The pathway is shown in Figure 3.

The enzymes 4ABH and NhoA were synthesized or PCRed from source material and cloned into a p15A vector with constitutive expression. The PabABC genes were assembled into an operon without a promoter and cloned into a high copy vector. As acetaminophen is a biologically active molecule in humans, we performed all experiments in *dapD* knockout strains that are unable to colonize the human gut.

The *in vivo* activity of the individual enzymes were confirmed by feeding in 10mM of precursors p-aminobenzoic acid or 4-aminophenol and using LCMS to confirm the 1-step transformations (p-aminobenzoic acid to 4-aminophenol and 4-aminophenol to acetaminophen, respectively). 4-aminophenol was detected via LCMS when p-aminobenzoic acid was fed to cells overexpressing 4ABH, and acetaminophen was detected via LCMS when 4-aminophenol was fed to cells overexpressing NhoA. The 2-step pathway was then confirmed by supplementing the media with 1 mM p-aminobenzoic acid and observing the production of acetaminophen. Finally, cells with the 4ABH, NhoA, and PabABC constructs were grown in glycerol minimal media to relieve product inhibition of the chorismate pathway and a yield of 2.9 μ M was measured via LCMS. No further optimization was attempted to improve the yield. The experimental confirmation of the acetaminophen pathway demonstrates how novel biosynthesis routes can be designed by computational algorithms and highlights the need for more formalizable biochemical data.

Methods

The Act Synthesizer

The Act Ontology is an aggregator of observations and the Act Synthesizer is a predictive tool that algorithmically explored the aggregated observations space. Each observation in Act is one biochemical interaction: binding, catalysis, dissociation etc. For MCF we are only concerned with catalysis observations. To compute all reachables we start with the set of native metabolites and iteratively do the following: Optimistically assume all enzymes that are *enabled*, i.e., whose substrates and all cofactors at this step are reachable, can be expressed in the host resulting in a new set of reachables as augmented the products of that catalysis. Doing this iterative wavefront expansion until no further enzymes are enabled yields the set of all reachables. We can then narrow this down by screening using substructure patterns (e.g., for reachable diesters, and diacids, etc.) or using enumerated datasets (e.g., DEA regulated chemicals, or known LD50 values, Drugbank small molecules, or Sigma-Aldrich functionally categorized chemicals) or by keyword (e.g., “cancer” for Drugbank entries that mention cancer in the associated text for the small molecule.) Supplement X contains examples of such screened reachables lists.

To generate the enzymatic cascade to a reachable chemical, the algorithm traverses backwards iteratively including all possible steps that could lead to the intermediate, ignoring at each step any enzymes that need an unnatural substrate. This condition ensures that it never includes a pathway step which lacks biosynthesizable precursors in the host. Once we have the entire space of possible pathways, i.e., the cascade, we then subsequently rank within them based on available expression data, and number of similar catalyses observed. These metrics assign a confidence score to each end-to-end pathway within the cascade.

This is a three phase algorithm: forward wavefront for all reachables (AND), backwards inclusion of pathway possibilities (OR), and ranking within the possibilities, and is a novel approach that exploits the semantics of enzymatic catalysis to develop an efficient algorithm that has previously eluded researchers. Previous attempts to solve this ranked hypergraph traversal problem (which is at least NP-hard) that defines the pathway construction process typically default to simulation-like incomplete solutions because of the hardness of the underlying computational problem. Act's phased solution not only provides valuable intermediate artifacts, i.e., all possible reachables, but is also more efficient.

Acetaminophen biosynthetic pathway

The *Agaricus bisporus* gene 4ABH was synthesized in-house using a LCR-based approach and the *E. coli* gene NhoA was cloned from genomic DNA. Both genes were placed under the control of the constitutive promoter BBa_J23100 in a p15A plasmid (4ABH.NhoA construct). The pabABC genes were assembled into an operon without a promoter and cloned into a high copy pUC vector (pabABC construct). As a biosafety precaution, we employed a *dapD* knockout strain that strictly requires diaminopimelic acid for growth. DapD *E. coli* knockout cells were generated via P1 transduction from the Keio collection [9]. DapD *E. coli* strains were transformed with either the 4ABH.NhoA construct or both the pabABC and 4ABH.NhoA constructs. Transformants were grown for two days in GMML (Teknova) supplemented with the corresponding antibiotics.

For LCMS analysis (using an Agilent 1260 Infinity HPLC and a Agilent 6120 Quadrupole LC/MS, using an Agilent Eclipse Plus C18 column with injection volumes 5 microliters), fermentation broth with cells were desalted using Waters Oasis HLB Light Cartridges for both the experimental and background samples. As acetaminophen is permeable to the *E. coli* membrane, we did not need to lyse cells and 1 mL of "crude" fermentation broth was loaded into the desalting cartridge. The negative control background was the Δ *dapD* strain transformed with an empty vector. The positive control was a negative control strain with acetaminophen (Sigma Chemicals) doped into the fermentation broth prior to desalting. A standard curve was generated for yield quantitation.

Discussion

As enzymatic knowledge increases, the space accessible by genetic engineering is also expected to increase. Computational predictions of the reachable space and their cascades is made possible by a formalization of biochemistry, as encoded by the Act Ontology and by algorithms that process it. Such formalizations and integration across different data sets (drug bioactivity data, material physiochemical data, flavor and fragrance profiles, etc.) allows for exhaustive enumeration of the opportunities and limits of MCF applications. Given the incomplete standardization of enzymatic, chemical, and cellular chassis data computational tools are in a state of continuous improvement. However, as we show with Act's characterization of the L0-L2 space as 3519 reachables, standardization of biochemical data is possible and useful for MCF engineering. We have also demonstrated the utility of our algorithms by experimentally engineering *E. coli* to produce a non-natural chemical, acetaminophen. Yet more data exists in published form. With ongoing future work on text mining, prediction of enzymatic

catalysis and binding, protein engineering models, virtual screening, and multiplex characterization, more data will be formalized into Act and will unlock an expanded space of reachable chemicals.

References

1. Paddon, C. J., et al. "High-level semi-synthetic production of the potent antimalarial artemisinin." *Nature* (2013).
2. Yim, Harry, et al. "Metabolic engineering of *Escherichia coli* for direct production of 1, 4-butanediol." *Nature chemical biology* 7.7 (2011): 445-452.
3. Berry, David A. "Engineering organisms for industrial fuel production." *Bioengineered* 1.5 (2010): 303-308.
4. Hatzimanikatis, Vassily, et al. "Exploring the diversity of complex metabolic networks." *Bioinformatics* 21.8 (2005): 1603-1609.
5. Moriya, Yuki, et al. "PathPred: an enzyme-catalyzed metabolic pathway prediction server." *Nucleic acids research* 38.suppl 2 (2010): W138-W143.
6. Pramanik, J., and J. D. Keasling. "Stoichiometric model of *Escherichia coli* metabolism: Incorporation of growth-rate dependent biomass composition and mechanistic energy requirements." *Biotechnology and bioengineering* 56.4 (1997): 398-421.
7. Schomburg, Ida, et al. "BRENDA in 2013: integrated reactions, kinetic data, enzyme function data, improved disease classification: new options and contents in BRENDA." *Nucleic acids research* 41.D1 (2013): D764-D772.
8. Yamamura, Ei-Tora, et al. "Purification and biochemical properties of an *N*-hydroxyarylamine *O*-acetyltransferase from *Escherichia coli*." *Biochimica et Biophysica Acta (BBA)-General Subjects* 1475.1 (2000): 10-16.
9. Baba, Tomoya, et al. "Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection." *Molecular systems biology* 2.1 (2006).

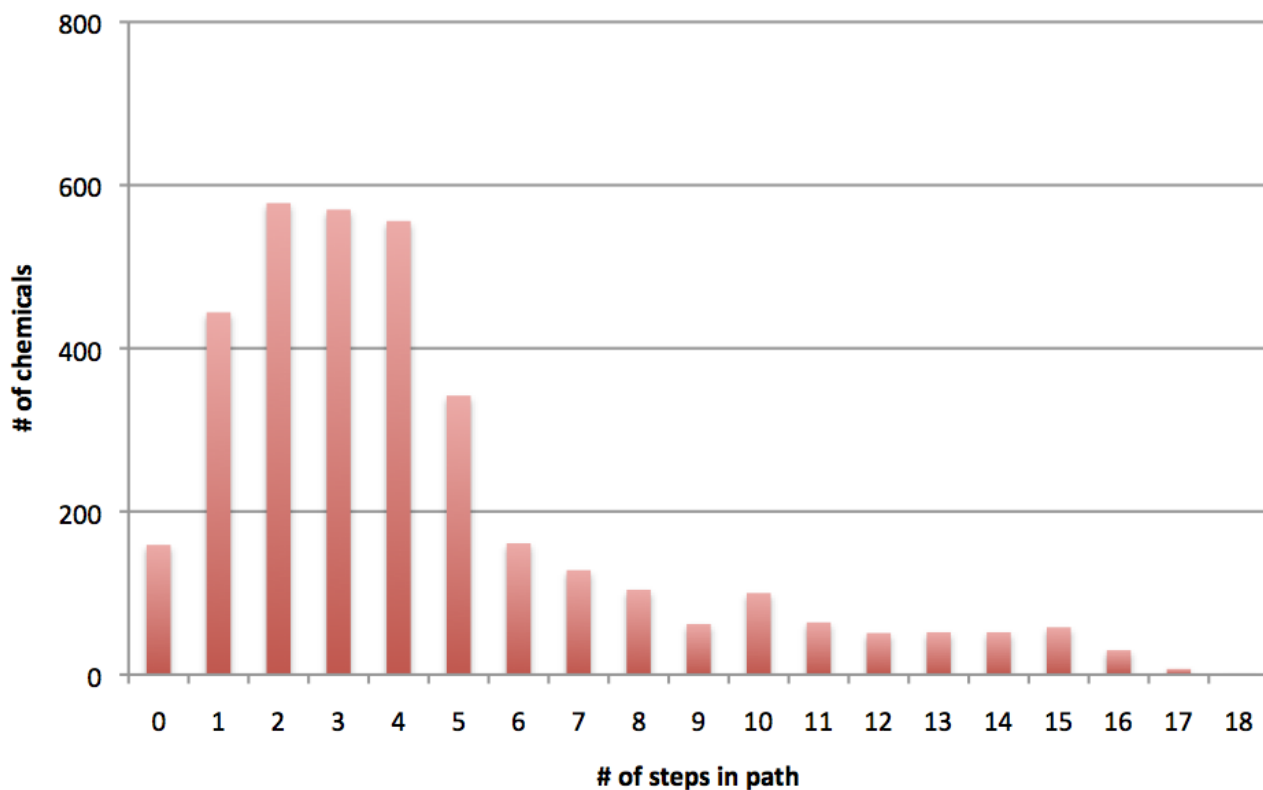


Figure 1. Graph of steps versus number of reachables. Graph showing how many steps from L0 (0 steps) are needed to enable reachables in the L0-L2 space. There are 3519 reachables in the L0-L2 space for *E. coli* using only data from BRENDA. This number increases with the addition of more biochemical observations from text mining (data not shown).

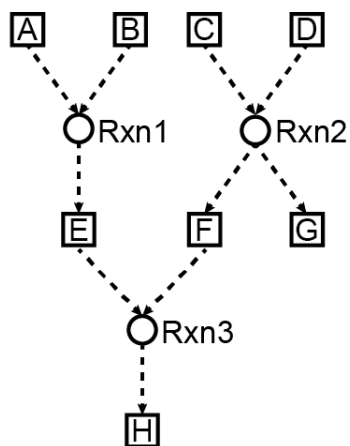
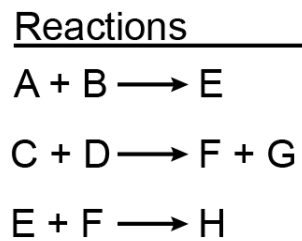


Figure 2. Bipartite representation of biochemical reactions. The three biochemical reactions on the left are represented by the bipartite, directed graph on the right. Squares are chemicals and reactions are circles. Manually defined cofactors such as NADH, ATP, etc. are not shown in this representation but are accounted for in the algorithm.

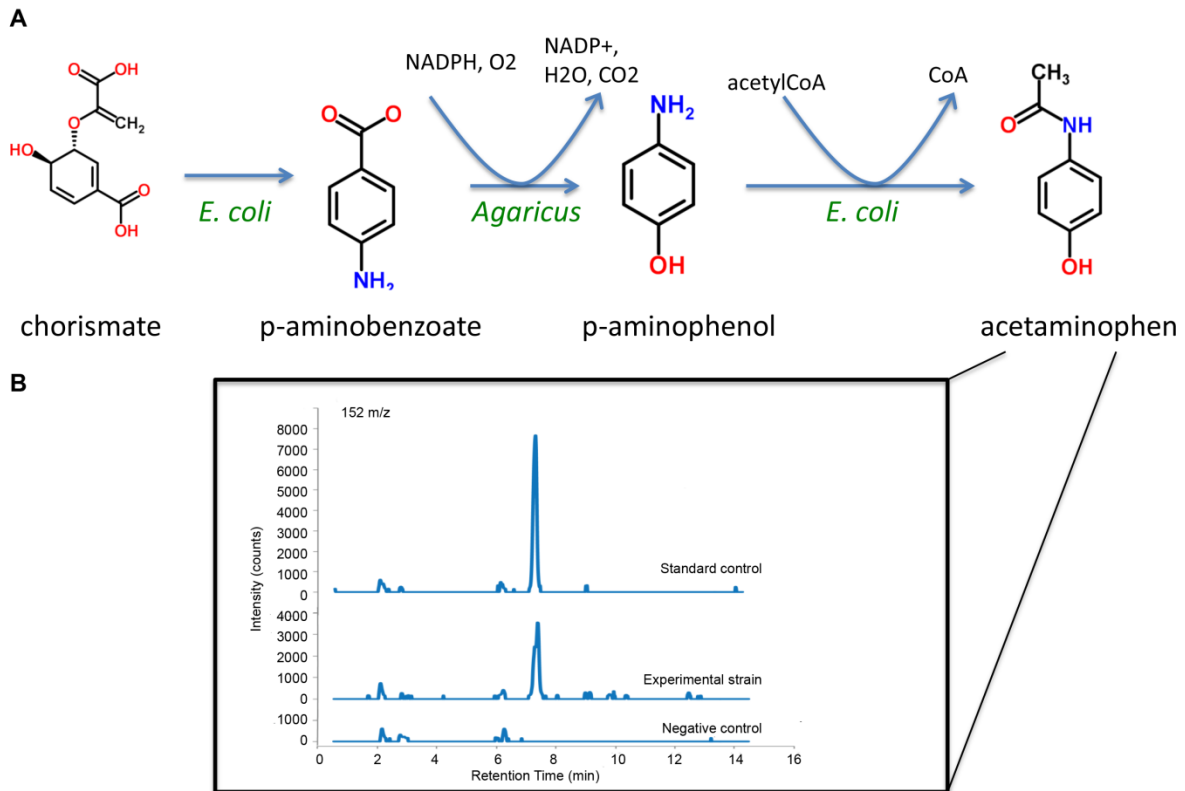


Figure 3. Demonstration of acetaminophen production in *E. coli*. A) depicts the Act-designed enzymatic pathway with organism sources for each enzymatic step. The LCMS trace is shown in B). Negative control strains were the $\Delta dapD$ strain transformed with an empty vector. Standard control strains were the negative control strain with acetaminophen supplemented in the growth media. The experimental strain expressed the 4ABH (from *Agaricus*) and the NhoA genes.

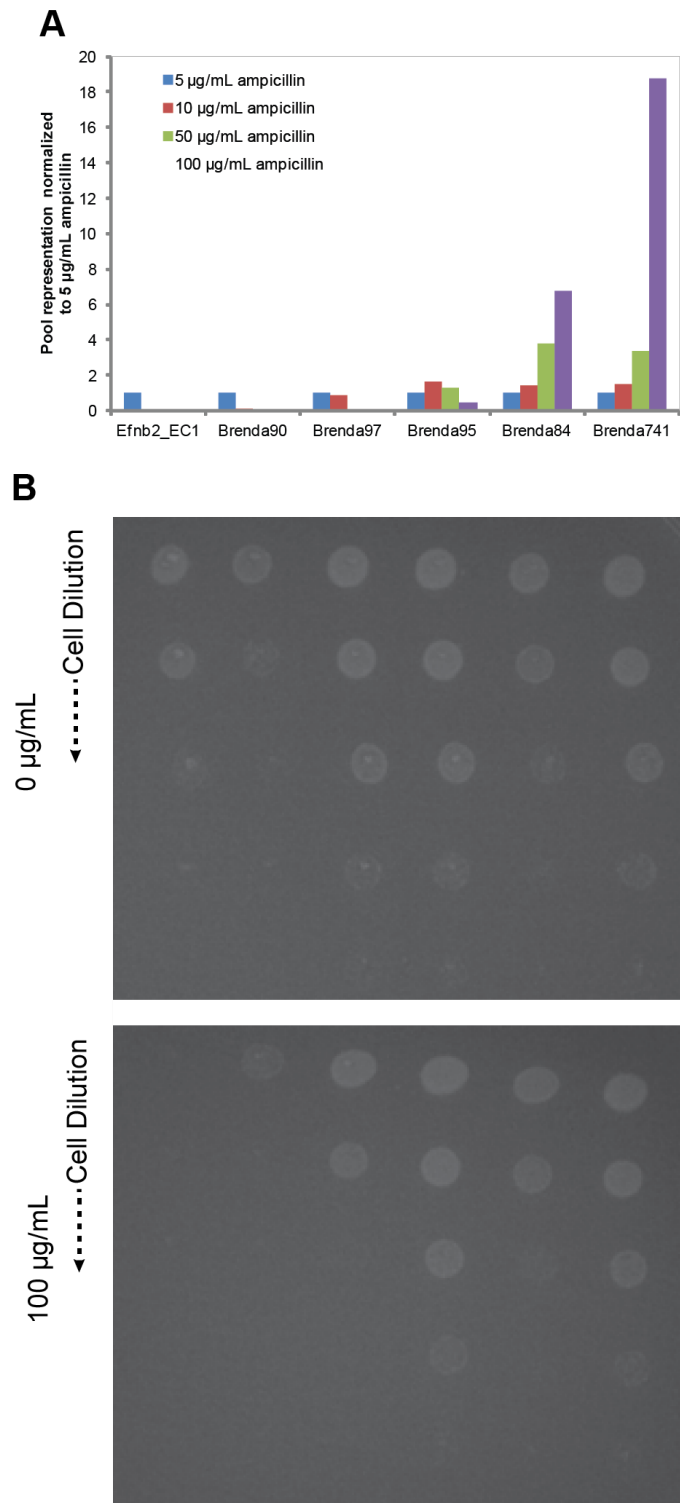


Figure 3. Multiplex sequencing correlates with plate expression phenotype. A) Graph of normalized occurrence of six representative ORFs. See the Methods for details on data normalization. B) Growth phenotype of strains on plates with different amounts of ampicillin.

Chapter 4

Conclusions

Genetic engineering has much promise but still remains a difficult endeavor. Biological systems are highly complex and attempts to modularize or standardize with the goal of predictive design have been met with limited success. Current state-of-the-art techniques all rely on some amount of undirected random sampling. However, there are certainly areas of biological engineering that are amenable to a more systematic approach. The results presented in this dissertation are a small suggestion of how a combination of empirical and predictive approaches can decrease the difficulty of genetic engineering.

Certainly, engineers and researchers will still grapple with the lack of predictability in biological research. Yet no technique will be able to exhaustively sample all genotype space for non-trivial systems, and most systems can only be assayed via laborious methods that only yield a few data points. However, as multiplex tools such as those described in this dissertation become widely available, I believe that the combination of better design methods, more data, and high-throughput methods of testing hypotheses will unlock ever-larger application spaces for biological engineering.