# Fluxomic analysis-based approach for engineering an ethanologenic strain of *Haloferax volcanii*

Josh L. Espinoza

**Principle Investigator:** David Bernick
jolespin@ucsc.edu
Submitted: May 7, 2016
Resubmitted: June 3, 2016

*Biomolecular Engineering*
*Jack Baskin School of Engineering*
*University of California, Santa Cruz*

# Abstract

Plant biomass represents a renewable carbon resource that could potentially replace fossil fuel based energy. A major obstacle impeding the integration of biofuels into the competitive market is the difficulty in gaining enzymatic access to the carbon-rich fermentable sugars embedded within the cell-wall's recalcitrant structural polymers. Partitioning lignin from cellulose is a fundamental step in the production process and typically requires introducing the biomass to ionic liquids; environments unsuitable for most lignocellulolytic and fermentation enzymes. Extremophile evolution has generated highly robust organisms that can tolerate the demands of industrial conditions; in particular, halophiles. While no halophile known to science efficiently synthesizes ethanol, advancements in computational biology have introduced novel approaches to direct carbon flux and to create a driving force mechanism to increase ethanol production. This research presents a pipeline for synthesizing a minimal ethanologenic strain of *Haloferax volcanii* using elementary mode analysis for directing carbon flux in parallel with a NADH-dependent driving force to couple growth with ethanol production. Included in this manuscript is a suite of software tools for optimal synthetic sequence generation and pathway informatics along with ortholog and reaction tables for mapping metabolic pathways across species. Lastly, this work contains physical DNA components for knocking out and overexpressing genes in *H. volcanii*.



Tags: fluxomics, elementary mode analysis, biofuels, haloferax volcanii, synthetic biology


# 0.0. Preface

## 0.1. Fossil fuels are unsustainable

Fossil fuels are combustible geological deposits formed over millions of years deep within the Earth's crust from the remains of living organisms. These petroleum-based fossil fuels supply ~82% of the energy demands in the U.S. and these resources are far from sustainable (Center for Sustainable Systems, 2015); the quest for renewable energy is of the utmost importance. One seamless alternative to fossil fuels, for use in combustion engines, is bioethanol; a low-emission biofuel that has an energy density (19.6 MJ/L) roughly comparable to gasoline (32 MJ/L) and can be used in unmodified gasoline engines (Cheng, 2009). Bioethanol can be synthesized directly from pentose and hexose sugars in cellulose by virtue of microbial organisms (Nakayama, Kiyoshi, Kadokura, & Nakazato, 2011).

## 0.2. Plant biomass as a carbon-neutral renewable resource

Plant biomass (lignocellulosic biomass) represents a renewable carbon source that could potentially replace fossil fuel based energy. Lignocellulosic biomass is the most abundantly available raw material on Earth for the production of biofuels (Carroll & Somerville, 2009). Lignocellulose is composed of carbohydrate polymers (i.e. cellulose and hemicellulose) along with the aromatic polymer lignin. Cellulose fibers are linked together with lignin polymers providing the strength and rigidity of plant cell walls. As lignin represents up to 30% of lignocellulosic material, it remains a major obstacle in biofuel production by impeding enzymatic access to the carbon-rich fermentable sugars embedded within the cell-wall's recalcitrant structural polymers (Brown & Chang, 2014). Fermentable sugars can be extracted from lignocellulosic biomass by decomposing the homopolymer cellulose into hexoses and the heteropolymer hemicellulose into both hexoses and pentoses (Kim, Block, & Mills, 2010). Hexoses and pentoses can be decomposed to pyruvate in the glycolysis pathway and utilized by fermentation pathways. In order to

break down the lignocellulosic polymers into fermentable sugars, lignin must first be separated from cellulose; a fundamental step in the biofuel production process.

One method for partitioning cellulose and lignin is by using high ionic strength organic solvents known as Ionic Liquids with temperatures ranging from 45 - 99°C, pH between 4.0 and 8.0, and concentrations up to 40% (Gladden et al., 2014). Most enzymes that are capable of processing cellulose and catalyzing fermentation do not function in these high ionic environments. This poses a problem because the layers of desalination steps required to reduce the ionic strength of the conditions require energy, time, and money. However, the infinite wisdom of Nature has provided all the materials and cellular machinery necessary for solving this problem. This evolutionary wisdom has generated organisms known as Halophiles that can thrive in extremely high ionic and saline environments (≥ 30%) when compared to the ocean (~3.5%) (Ollivier, Caumette, Garcia, & Mah, 1994).

## 0.3. Ethanol Fermentation Pathways

The most efficient metabolic reactions for producing ethanol require the catalysis of pyruvate substrate into acetaldehyde followed by acetaldehyde substrate into ethanol via pyruvate decarboxylase and alcohol dehydrogenase, respectively (Trinh, Unrean, & Srienc, 2008). Integrating pyruvate decarboxylase into the metabolism mimics the ethanol-producing pathways of native ethanologenic organisms such as *Zymomonas mobilis* and *Saccharomyces cerevisiae*. Although conversion of pyruvate to acetaldehyde can be mediated via pyruvate formate lyase with an acetyl-CoA intermediary (highly regulated and, thus, rate limiting), the affinity of pyruvate to pyruvate carboxylase is much higher than that of pyruvate formate lyase and is not dependent on acetyl-CoA (Ingram & Conway, 1988); an ideal shunt for a minimalist ethanologenic design as shown in Figure I.

Formation of lactate and malate consume valuable carbon resources (i.e. pyruvate) and NADH, a high valued redox-related currency, limiting the amount of substrate for producing ethanol product. NADH is oxidized not only in the conversion of acetaldehyde to ethanol but also in competing pathways converting pyruvate to lactate or succinate. Therefore, maximizing the pool of reduced NADH for use by alcohol dehydrogenase and the simultaneous synthesis of ethanol with NAD+ recycling. Table I shows the efficiency of ethanol fermentation reactions in *E. coli* using hexoses and pentoses as substrates in terms of mass ratios. Two of the elementary modes (refer to 0.5. for details on elementary modes) represented by the reactions in Table I coproduce biomass and ethanol during growth phase with a productive ethanol yield of 0.45 g ethanol/g of sugars. Thus, implying that a recombinant strain with the binary purpose of growth and ethanol production could be engineered to utilize these pathways. These reactions are elaborated on in sections 2.1. and 3.1.

.

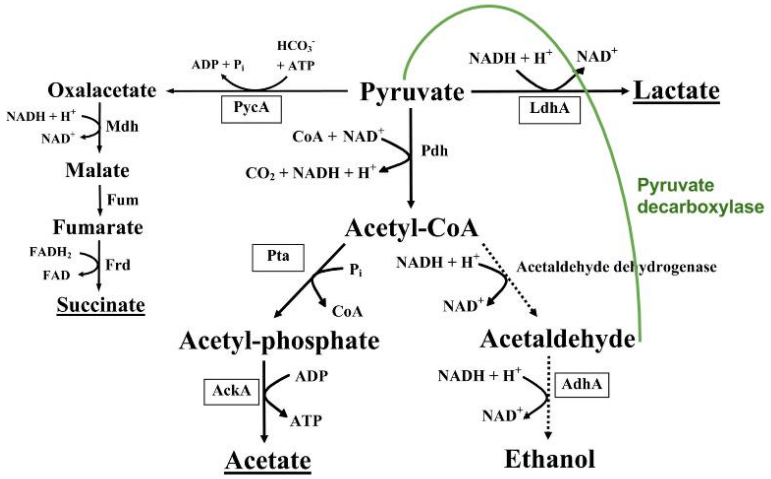


**Figure I I** Pyruvate fate in varying metabolic reactions
The fate of pyruvate and various pathways it can traverse. NADH reduction in malate and lactate reactions deplete resources from acetaldehyde conversion into ethanol. In green, the shunt introduced by pyruvate decarboxylase providing a direct route from pyruvate to acetaldehyde by passing dependency on acetyl-CoA.
Figure source: (Eschbach et al., 2004)

**Table I I** Stoichiometric equations for efficient ethanol-producing reactions
Stoichiometric reactions showing elementary modes that produce ethanol; 2 of which couple growth to ethanol production.
Table source: (Trinh et al., 2008)

| Substrate | Equation | $Y_{ETOH/sugar}$ (g/g) |
|---|---|---|
| Glucose | 1 glucose = 2 ethanol + 2 $CO_2$ | 0.51 |
| | 1 glucose = 2 ethanol + 2 $CO_2$ + 2 ATP | 0.51 |
| | 1 glucose + 0.49 $NH_3$ = 1.79 biomass + 1.40 ethanol + 0.11 $H_2$ + 1.41 $CO_2$ | 0.36 |
| | 1 glucose + 0.49 $NH_3$ = 1.79 biomass + 1.40 ethanol + 0.11 formate + 1.30 $CO_2$ | 0.36 |
| Xylose | 1 xylose = 1.67 ethanol + 1.67 $CO_2$ | 0.51 |
| | 1 xylose = 1.67 ethanol + 1.67 $CO_2$ + 0.67 ATP | 0.51 |
| | 1 xylose + 0.16 $NH_3$ = 0.60 biomass + 1.47 ethanol + 0.04 $H_2$ + 1.47 $CO_2$ | 0.45 |
| | 1 xylose + 0.16 $NH_3$ = 0.60 biomass + 1.47 ethanol + 0.04 formate + 1.43 $CO_2$ | 0.45 |

[a] The measure of biomass is 1 Cmol.

## 0.4. Halophiles for Biotechnology

The industrial usage of mesophilic enzymes as biocatalysts is constrained to a narrow bandwidth of activity ranges for pH, pressure, salinity, and temperature. Haloarchael proteins, often described as extremozymes, have evolved to remain soluble, stable, and catalytically active in high ionic conditions (Coquelle et al., 2010) including both hypersaline and organic solvent environments (Flam, 1994). With these unique characteristics, these specialized proteins tend to tolerate extreme conditions when compared to their mesophilic orthologs (Timpson et al., 2012). Halophilic extremozymes are able to keep their proteins from precipitating in high ionic strength conditions by incorporating an abundance of acidic amino acid residues while containing most of the negative charges on the surface; helping to orient water molecules and sustain hydration (Kennedy, Ng, Salzberg, Hood, & DasSarma, 2001).

Halophilic archaea are prime candidates for biotechnological applications not only for their robust activity,

but also their basic architecture and ability to generate large quantities of biomass. Many haloarchaea require low maintenance culturing protocols and are not subject to contamination due to the high salt concentrations in which they thrive. Haloarchaea contain unique features that make them attractive targets for industrial use such as lysis of the cell in low salt environments, thus, reducing the costs of protein purification. The most attractive feature of haloarchaea, through the lens of biofuel production, is the aforementioned high salinity adapted enzymatic activity. *Haloferax volcanii* is an extremely halophilic polyextremophile and is a model organism for the Archaea domain (Margesin & Schinner, 2001); a promising microbial target for engineering biofuels.

## 0.5. Elementary Mode Analysis

Microbes have evolved to withstand a diverse range of environmental conditions allowing them to adapt and colonize a variety of ecosystems. This ability is largely due to the redundancy of metabolic pathways. Their modular metabolism aids in facilitating growth utilizing a variety of input substrates and tolerance to environmental fluctuations. A metabolic pathway is a series of chemical reactions occurring within a cell and, due to the redundancy of adaptable evolution, can be decomposed into metabolic building blocks known as Elementary Modes [EMs] (Trinh, Wlaschin, & Srienc, 2009). EMs are defined as non-decomposable steady-state pathways (i.e. concentrations remain constant with respect to time) through a metabolic network. The non-decomposability aspect entails that if any of an EM's composite reactions are not present, the EM will not be able to maintain a steady-state (Zanghellini, Ruckerbauer, Hanscho, & Jungreuthmayer, 2013). If a reaction is deleted, the function of the pathway will no longer be accessible for the metabolic system; minimal functional building blocks. EM Analysis is a tool that determines all possible solutions of a metabolic matrix by calculating a unique sets of elementary flux modes. Overlaying all of the elementary modes builds a complete metabolic network, a fluxome, for an organism.

Formally, a flux vector $\boldsymbol{v}$ in a metabolic network (represented as $\mathbf{S}$, a stoichiometry matrix), is a flux mode *iff*: (i): $\boldsymbol{v} \neq 0$; (ii): $\boldsymbol{v} \in R^n$ ; (iii) $\mathbf{S} \cdot \boldsymbol{v} = 0$ (i.e. solves steady-state condition); and (iv) follows (ir)reversibility constraints of reaction $\boldsymbol{v}$. The support of a flux mode, $\mathbf{supp}(\boldsymbol{v}) = \{ i \mid v_i \neq 0\}$ as the set of non-zero indices in the flux. A flux mode is an EM, $\boldsymbol{e}$, if $\mathbf{supp}(\boldsymbol{e}) \not\supset \mathbf{supp}(\boldsymbol{v})$ (i.e. $\boldsymbol{e}$ is not contained in a superset of any other $\boldsymbol{v}$) (Zanghellini et al., 2013).

Metabolic networks contain internal metabolites (assumed to be in steady-state) and external metabolites (acting as sources and sinks). The topology of these networks is described by $\mathbf{S}$, a $n \times m$ stoichiometry matrix, composed of $n$ reactions and $m$ internal metabolites. The scalar at $S_{i,j}$ represents the stoichiometric coefficient of metabolite $i$ in reaction $j$: $S_{i,j} > 0$ represents metabolite formation; $S_{i,j} < 0$ represents metabolite consumption; and $S_{i,j} = 0$ indicates the metabolite does not participate in reaction $j$ (Zanghellini et al., 2013). EMs are calculated by consecutively deleting a single reaction and solving for the steady-state condition until all remaining reactions are required to maintain the condition (Acuña et al., 2009). The composite nature of EMs and the modularity of metabolic networks requires that a complete set of EMs fully defines all pathways within a metabolic network.

Essentially, EM analysis identifies a unique set of flux modes that operate at steady-state. When EMs are identified, they can be used to optimize engineering strategies by creating minimal functionalities and identifying robustness of specific metabolic functions. Unfortunately, EM analysis is vastly demanding regarding compute resources due to the combinatorial explosion of EMs with increasing metabolic network size. To illustrate this, consider that a model for the core metabolism of *E. coli* contains ~100 reactions and consists of ~272 million EMs (Jungreuthmayer & Zanghellini, 2012). Current efforts for

efficiently computing EMs discussed in section 5.1.

# 1.0. Introduction

The prospect of biofuels not only addresses the moral dilemma of our environmental impacts as a species but also presents an opportunity to progress interdisciplinary science with economic incentive. As the price for gasoline fluctuates and the geological deposits of fossil fuels rapidly depleting, the switch to a compatible renewal resource is a logical transition. Considering the advancements in computational science used in parallel with cutting-edge systems- and synthetic-biology techniques, this biofuel-based future is within grasp for our generation.

The tools and documents presented in this manuscript (3.3.1. – 3.3.2.) will be useful for synthesizing recombinant constructs and completing various steps in metabolic engineering. The pathway and SBML files accompanying this text are manually curated exploiting the reproducibility of KEGG notation and can be used for cellular modeling with additional fluxomics analyses. The ortholog table is meant for identifying novel genes with similar protein domains as the knockout targets discovered by Trinh et al. 2008. The DNA components synthesized in 3.3.3. are intended for researchers conducting empirical experiments concerning biofuel-based metabolic engineering of *H. volcanii.*

Section 2.0. discusses previous research on individual components of this manuscript. Sections 2.1. and 2.2. focus on using pyruvate decarboxylase from *Zymomonas mobilis* for ethanol production. Section 2.1. takes an EM analysis based approach to direct carbon flux towards ethanol production in *E. coli* while 2.2. takes a biochemical approach and characterizes pyruvate decarboxylase in *E. coli* and *H. volcanii*. Section 2.3. measures relative expression in varying pH condtions and discovers that alcohol dehydrogenase II is upregulated in *H. volcanii* under acidic conditions; relating directly to 2.2. and their inability to catalyze the ethanol reaction. Section 2.4. compares various properties of alcohol dehydrogenase I & II in *H. volcanii* where they determine that alcohol dehydrogenase II is more robust than its paralog and an ideal candidate for exploitation by biotechnology. Lastly for the relevant studies, Section 2.5. develops a NADH-dependent driving force technique in synthesizing butanol in *E. coli* by mimicking the metabolic architecture of *Clostridium* species.

In unison, these studies form a structure whose architecture suggests that it is indeed possible to synthesize an ethanologenic strain of *H. volcanii*. Section 3.0 discusses the research project and, 3.1. in particular, the logic for why the previous studies point towards the direction taken by this body of work and how to build on past efforts by current literature. Section 3.3. details my project aims and specific goals that naturally arise from my logic and hypotheses in sections 3.1. and 3.2., respectively. Section 3.4. describes the approach for completing my aims and the software tools I used to create resources for future researchers studying fluxomics.

My research methods in Section 4.0. are presented in great detail where 4.1. - 4.4. focus on my computational approaches and 4.5. – 4.7. describe my molecular-biology methods to generate physical DNA components. Section 5.0. embeds a discussion that describes my findings and production materials in the context of future researchers. Finally, section 6.0. discusses current technology in fluxomics, biofuels, and synthetic biology along with the direction each field is heading in the future. I discuss this section in the framework of my interests and my future goals as an interdisciplinary scientist.

# 2.0. Relevant Studies

## 2.1. EM Analysis optimization of carbon flux for microbial ethanol production (Trinh et al., 2008)

The existence of inefficient pathways for ethanol production can lower the yield. These inefficiencies are largely to due to competing pathways using similar substrates. An example of decreased efficiency can be found in the (Dumsday, Zhou, Yaqin, Stanley, & Pamment, 1999) study that used an *E. coli KO11* chemostat culture. This study reported that growth on pentose is diminished as *KO11* loses its hyperethanologenicity at the expense of cell growth and acetate synthesis. Trinh et al. 2008 addressed the importance of metabolite flux in ethanol pathways by eliminating non-essential ethanologenic substrate sinks and non-directed carbon flux.

Pyruvate, a key substrate for ethanol production, is a promiscuous substrate that interacts with several metabolic pathways and is an ideal candidate for optimization during biofuel production. Trinh et al. 2008 expressed foreign pyruvate decarboxylase (*Zymomonas mobilis)* to introduce a direct route from pyruvate to acetaldehyde. Using EM analysis, Trinh et al. 2008 identified non-essential pyruvate sinks in *E. coli* and found that perturbing these pathways, listed in Table II, provides more pyruvate substrate for ethanol production (Figure II); increasing the range of ethanol yield to 0.36 - 0.51 g ethanol/g sugars. The disruption of competing pyruvate pathways directed carbon flux towards the exogenous pyruvate decarboxylase; one source of this increase in yield.

Based on the *E. coli* metabolic network, Trinh et al. 2008 identified 15,185 EMs that metabolize pentoses. Of the pentose utilizing EMs, 1,004 EMs function in anaerobic conditions. Of the anaerobic EMs, 964 produce ethanol, 443 synthesize biomass, and 415 produce both ethanol and biomass. EM analysis identified that a set of 7 reactions (encoded in 8 genes displayed in Table I) reduced the dimensionality of the EM space significantly from 1,004 to 12. Of the 12 EMs, all 12 produce ethanol and 4 of which support growth. Note that the reduction of EMs displayed in Figure III are positively correlated with increasing ethanol yield. As previously mentioned, pyruvate formate lyase also uses pyruvate as a substrate to produce ethanol but with low yield. Unfortunately, Trinh et al. 2008 found that deletion of this enzyme prevents growth under anaerobic conditions because the synthesis of acetyl-CoA is required for producing biomass.

**Table II I** Ortholog table conversion from *E. coli* to *H. volcanii*
The genes targeted for knockout out in *E. coli* by Trinh et al. 2008 with their respective pathway labels and *H. volcanii* orthologs.

| Identifier | Symbol | Gene | Pathway | *E. Coli* | *H. volcanii* |
|---|---|---|---|---|---|
| **ANA2** | sfcA | Malic enzyme A | b1479 | P26616 | L9V927 |
| **ANA2** | maeB | Malic enzyme B | b2463 | P76558 | D4GV91 |
| **FEM3** | *ldhA* | Lactate dehydrogenase | b1380 | P52643 | D4GZF9 |
| **FEM7** | *pta* | Phosphate acetyltransferase | b2297 | P0A9M8 | NaN |
| **PPP1** | *zwf* | G6P-1 Dehydrogenase | b1852 | P0AC53 | D4GS48 |
| **OPM4r** | *ndh* | NADH dehydrogenase | b1109 | P00393 | D4GVG0 |
| **FEM2** | *poxB* | Pyruvate dehydrogenase | b0871 | P07003 | HVO_0669 |
| **TCA10** | *frdA* | Fumarate reductase | b4154 | P00363 | D4GX38 |

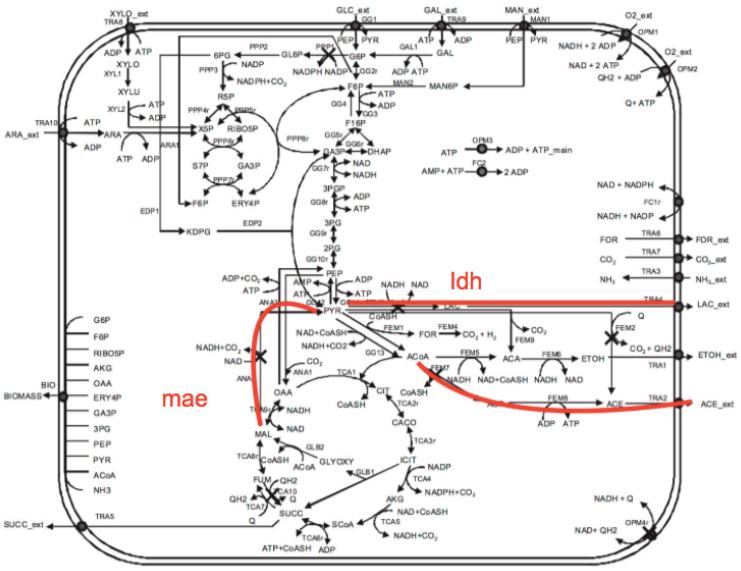


**Figure II** I Metabolic map of EMs in fermentation network

Listed are glucose-6-phosphate-1-dehydrogenase (PPP1, zwf), NADH dehydrogenase II (OPM4r, ndh), NAD/NADP-dependent malate enzyme (ANA2, sfcA/maeB), D-lactate dehydrogenase (FEM3, ldhA), fumarate reductase (TCA10, frdA), pyruvate oxidase (FEM2, poxB), and phosphate acetyl- transferase (FEM7, pta). ETOH, ethanol. My project seeks to disrupt ANA2 and FEM3 pathways labeled in red as mae and ldh, respectively.

Figure source: (Trinh et al., 2008)

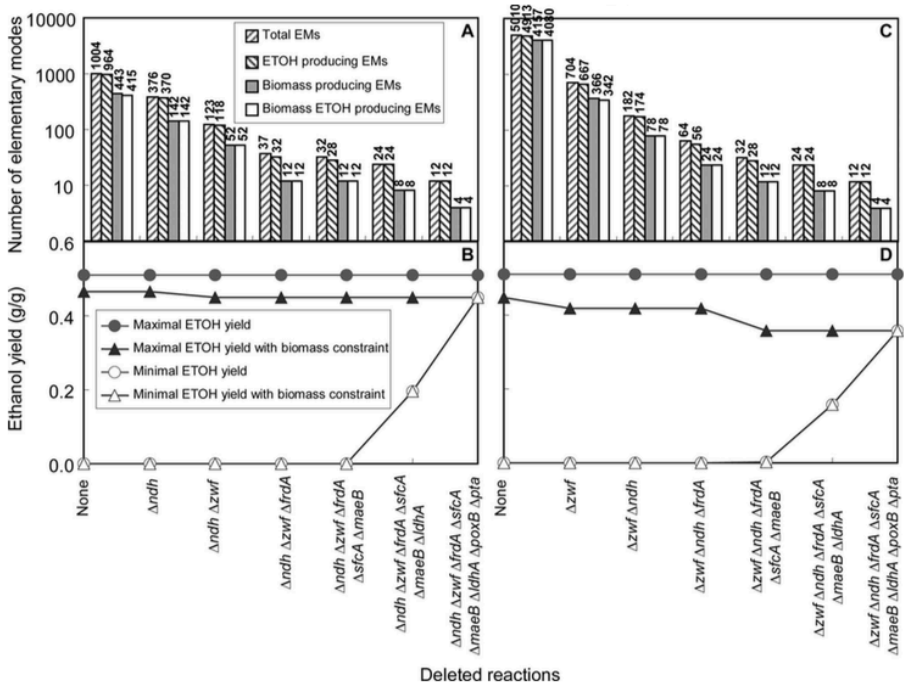


**Figure III** I Ethanol yield with varying pathway disruptions

Number of anaerobic EMs for growth on pentoses (A & B) and hexoses (C & D) in relation to ethanol yield (ETOH). In each group of bars, the numbers of: (i) total modes; (ii) modes that make ethanol; (iii) modes that produce biomass; and (iv) modes that make both biomass and ethanol are listed.

Figure source: (Trinh et al., 2008)

## 2.2. Pyruvate decarboxylase expression in *H. volcanii* (Kaczowka, Reuter, Talarico, & Maupin-Furlow, 2005)

The Maupin-Furlow group recognized the potential of halophilic archaeon in biotechnological applications and attempted to produce ethanol using *H. volcanii* (Kaczowka et al., 2005). In this study, Kaczowka et al. 2005 employed pyruvate decarboxylase based on its application in the industrial production of biofuels (Iding, Siegert, Mesch, & Pohl, 1998). Hoogenraad's group found that *Zymomonas mobilis* pyruvate decarboxylase [ZmPDC] purified from recombinant *E. coli* is comparable with native ZmPDC with respect to biochemical characteristics (Neale, Scopes, Wettenhall, & Hoogenraad, 1987). Ingram's group revealed that ZmPDC is stable and active at high temperatures (50 – 60 ˚C) (Ingram & Conway, 1988). These findings indicate that ZmPDC is an ideal candidate for biofuel applications using *H. volcanii.*

Kaczowka et al. 2005 proposed that ZmPDC may be stabilized in high salt concentrations by virtue of the same structural forces responsible for high thermal stability. To investigate their hypothesis, they expressed and purified ZmPDC from recombinant *E. coli* and assayed the activity at increasing salt concentrations as displayed in Figure IVA. The results indicate that ZmPDC retained roughly 50% of its activity at 2 M salt while retaining 20% activity at 4 M salt. This finding was used to justify the use of ZmPDC as a direct channel from pyruvate to acetaldehyde in the high ionic environment of *H. volcanii*'s cytosol. The presence of ZmPDC was confirmed via immunoanalysis on log-phase recombinant *H. volcanii* cells using anti-ZmPDC antibody shown in Figure IVB. In regards to PDC activity, a low level of ZmPDC activity (0.12 U per mg protein) was present in the recombinant *H. volcanii* cell lysate but null in the parent strain.

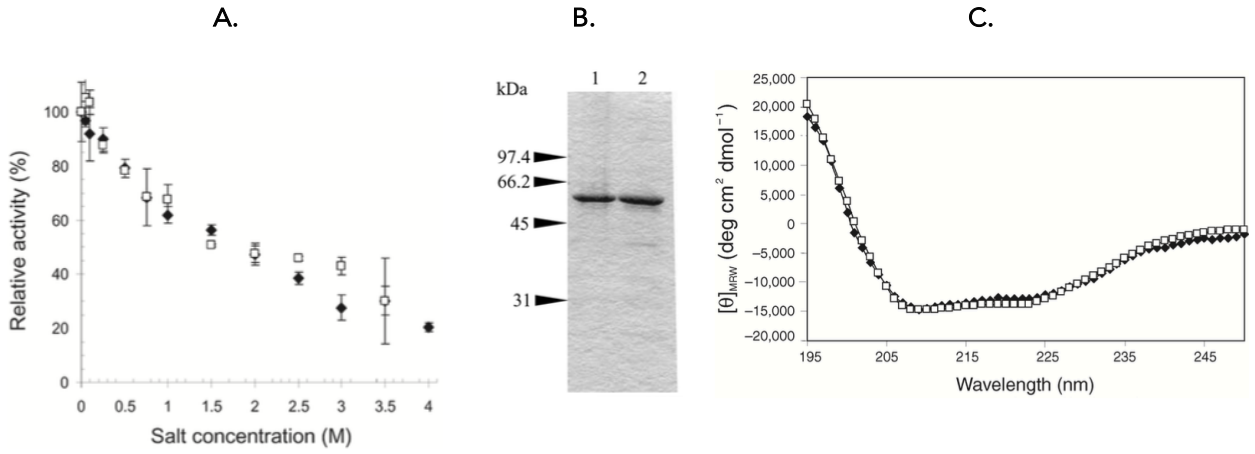


**Figure IV I** *Z. mobilis* pyruvate decarboxylase characteristics in recombinant microbes
(A) Effect of salt on the pyruvate decarboxylase activity of *Z. mobilis* purified from recombinant *E. coli;*
(B) Comparison of ZmPDC purified from recombinant H. volcanii (Lane 1) with that from recombinant E. coli (Lane 2);
(C) Circular dichroism of ZmPDC purified from *H. volcanii and E. coli* spectra at 100 – 160 nM (5mM sodium phosphate buffer, pH 6.5)
Figure source: (Kaczowka et al., 2005)

## 2.3. Differential gene expression in *H. volcanii* under extreme pH (Moran-Reyna & Coker, 2014)

As mentioned previously, haloarchaea are highly robust organisms that can withstand adverse

environmental conditions such as elevated temperatures and high ionic environments (Kaczowka et al., 2005) (Coker, DasSarma, Kumar, Müller, & DasSarma, 2007). As much is known regarding haloarchaea's adaptability to varying levels of salt and temperature, studies have shown that these organisms have high tolerance to gamma radiation (DeVeaux et al., 2007), UV (Baliga et al., 2004), pH (Moran-Reyna & Coker, 2014), and oxygen levels (Müller & DasSarma, 2005).

The Coker group recognized these adaptations and investigated the differential expression of *H. volcanii* genes in response to acidic and alkaline conditions (6.0 and 8.0, respectively) while assaying growth curves between pH 4.5 – pH 8.5 as displayed in Figure IV (Moran-Reyna & Coker, 2014). Moran-Reyna et al. 2014 identified (869 upregulated | 987 downregulated) genes at pH 6.0 and (326 upregulated | 1581 downregulated) genes at pH 8.0 in *H. volcanii* when compared to the optimal pH of 7.5. The acidic findings show interesting results for the prospect of biofuel production; reasons described in section 3.1.

The upregulated and downregulated transcripts were composed of several groups, the largest of which played roles in amino acid transport and metabolism while the lowest percent functioned in RNA processing/modification and chromatin structure. While 33% and 46% of upregulated and downregulated transcripts, respectively, were of unknown function, the vast majority of upregulated transcripts coded for products involved with stress and motility such as *cct*2, *cct*3, *cspD*3, *cheDFR*, *flaD2J*, and *arcR*1. The response to acidic pH resembles the effects seen for *E. coli (Moran-Reyna & Coker, 2014).*

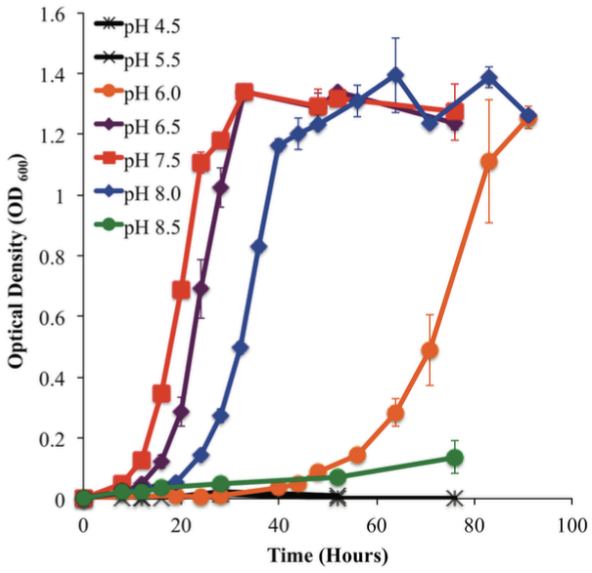


**Figure V** | Growth Curve of *H. volcanii* cultures grown under extreme pH conditions
Growth of *H. volcanii* at pH 4.5 (black, *), pH 5.5 (black, ×), pH 6.0 (orange, ●), pH 6.5 (purple, ♦), pH 7.5 (red, ■), pH 8.0 (blue, ♦), and pH 8.5 (green, ●). OD axis values for each point represents the average of all replicates and the error bars represent the standard deviation at each time point.
Figure source: (Moran-Reyna & Coker, 2014)

### 2.4. Alcohol dehydrogenase paralog characterization in *H. volcanii* (Timpson et al., 2013)

Alcohol dehydrogenases are important biocatalysts for interconversion of alcohols and aldehydes as shown in Figures I-II. For biotechnological applications of alcohol dehydrogenases, it is of vital importance that the enzyme can exhibit broad substrate specificity, thermostability, NADH dependence, and the

ability to tolerate high strength organic solvents.

The Paradisi group recognized the importance of these extremozymes and the lack of public knowledge available, thus, produced the first body of work on alcohol dehydrogenase paralogs from *H. volcanii* (Timpson et al., 2013). Their study focused on *H. volcanii* versions of *adh1* (HVO_2428) and *adh2* (HVO_B0071) genes whose protein products are referred to as HvADH1 and HvADH2, respectively. HvADH2 is the most thermoactive haloarchael alcohol dehydrogenase to date (Timpson et al., 2012). Preliminary data showed that both HvADH1 and HvADH2 catalyzed the redox reaction at an acidic pH (Timpson et al., 2013). As shown in Figure VI, HvADH2 has high activity for an expansive range of temperatures and shows high stability in both crude and pure isolates when compared to HvADH1. Timpson et al. 2013 also has shown that HvADH2 can tolerate organic solvents, acetonitrile and dimethyl sulfoxide, at concentrations of 10 and 20%, respectively. Timpson et al. 2013 provided evidence that HvADH2 accepted a broad range of substrates when compared to HvADH1; whose affinity spectrum was dominated by ethanol. Interestingly, HvADH2 had a higher affinity to acetaldehyde and NADH when compared to that of HvADH1 as shown in Table III. Given the extreme conditions of industrial use and stability-related enzyme costs, HvADH2 represents a broad substrate extremozyme capable of robust activity in high ionic and high temperature environments.

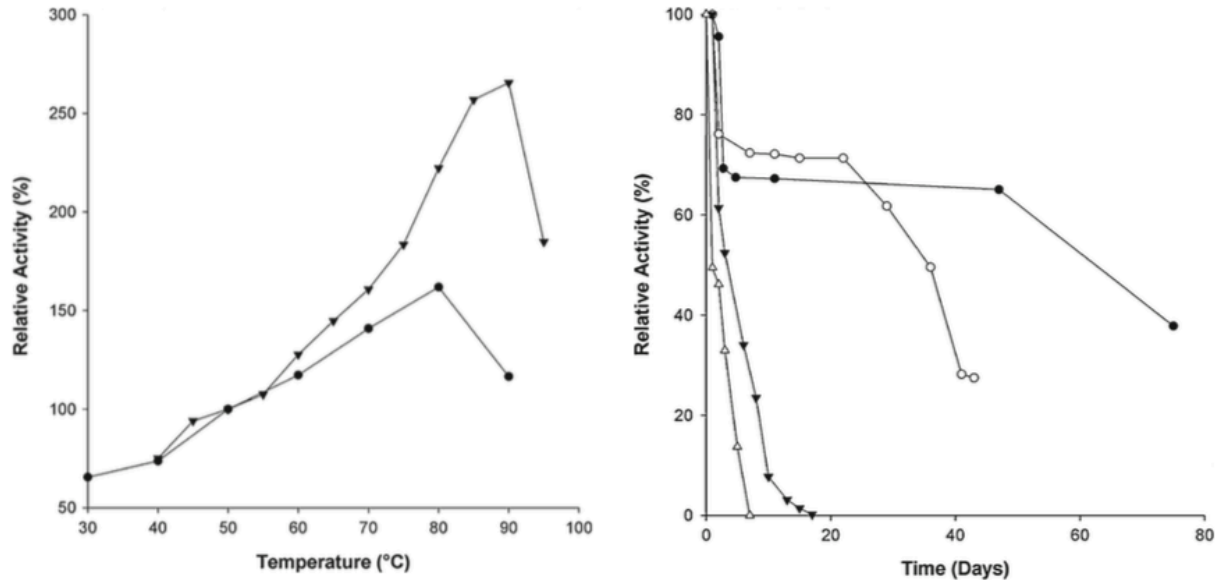


**Figure VI** | Thermoactivity and stability of *H. volcanii* paralogs
[Left] Thermoactivity of HvADH1 (●) and HvADH2 (▸) assayed for oxidative reactions at temperatures from 30 - 95 ˚C.
[Right] Stability of HvADH1 [(crude, ▹), (pure, ▸)] and HvADH2 [(crude, o), (pure, ●)] at -20 ˚C.
Results expressed as relative activities (%).
Figure source: (Timpson et al., 2013)

**Table III** | Kinetic parameters of HvADH1 and HvADH2

$K_m$ represents the Michaelis Constant of an enzyme which describes their tendencies to bind a specific substrate. A high $K_m$ correlates with a low affinity for the enzyme to bind a specific substrate and vice versa.

Table source: (Timpson et al., 2013)

| | *Hv*ADH1 | | | *Hv*ADH2 | | |
|---|---|---|---|---|---|---|
| Substrate/cofactor | $K_m$ (mM) | $V_{max}$ (μmol $min^{-1}$ $mg^{-1}$) | $K_{cat}/K_m$ ($s^{-1}$ $mM^{-1}$) | $K_m$ (mM) | $V_{max}$ (μmol $min^{-1}$ $mg^{-1}$) | $K_{cat}/K_m$ ($s^{-1}$ $mM^{-1}$) |
| Ethanol | 31±2 | 3.65±0.08 | 0.072 | 44±1 ($K_{0.5}$) | 5.01±0.07 | 0.070 ($K_{cat}/K_{0.5}$) |
| 1-Butanol | – | – | – | 6.7±0.2 ($K_{0.5}$) | 5.2±0.1 | 0.479 ($K_{cat}/K_{0.5}$) |
| Acetaldehyde | 28±3 | 2.2±0.1 | 0.048 | 9.2±0.9 | 11.1±0.3 | 0.744 |
| $NAD(P)^+$ | 0.29±0.02 | 4.28±0.08 | 9.037 | 0.21±0.01 | 4.22±0.09 | 12.390 |
| NAD(P)H | 0.071±0.008 | 3.8±0.2 | 32.771 | 0.032±0.002 | 12.6±0.4 | 242.763 |

## 2.5. "Driving Force" for optimizing redox reaction flux (Shen et al., 2011)

As fermentation can produce diverse products, the Liao group focus their efforts on optimizing the production of butanol in *E. coli* (Shen et al., 2011). Butanol is naturally synthesized by *Clostridium* species utilizing acetyl-coA as an intermediate. While not ethanol, the "Driving Force" method they have applied in this study provides an invaluable tool for the framework of synthetic biology and pathway optimization. Their high-yielding mechanism can be described by: (i) creating a driving force; and (ii) coupling the driving force to the desired pathway. Shen et al. 2011 identifies the mechanism that makes high-yield fermentation products possible, namely, a decarboxylation reaction near the end products causing an irreversible release of $CO_2$ that acts a driving force to pull fluxes towards a specific fermentation product. In *E. coli*, no significant driving force exists to direct the carbon flux starting from acetyl-CoA as in the clostridial CoA-dependent 1-butanol pathway (Atsumi et al., 2008).

The goal of their research is to delete competing fermentation pathways which increases the availability of NADH and to reengineer the 1-butanol pathway to utilize only NADH as the reducing co-factor; a driving force for butanol production in *E. coli* as shown in Figure VII. Introduction of *ter* from *Treponema denticola*, an irreversible NADH-dependent enzyme, facilitates a coupling of the desired fermentation pathway with the NADH-dependent driving force; producing a higher yield of product (Shen et al., 2011). By coupling the desired production pathway with the primary NAD+ regeneration route to growth, the reengineered organism may evolve greater catalytic efficiency and robustness.

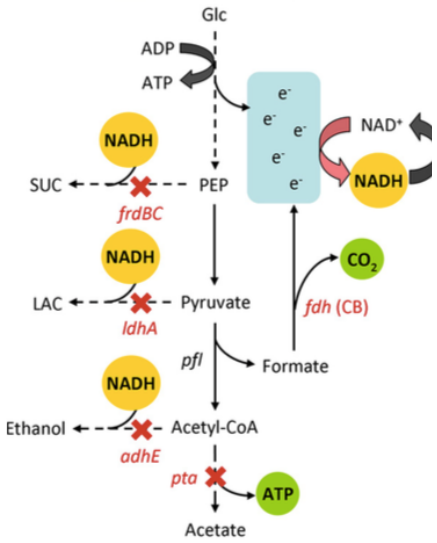


**Figure VII** | Synthetic build-up of NADH for increased reduction reactions
Pathways that utilize NADH as an electron donor are disrupted creating a surplus of reductive substrates (i.e. NADH)
Figure source: (Shen et al., 2011)

# 3.0. Master's Project

## 3.1. Logic

The Trinh et al. 2008 study showed that exogenous expression of the *Z. mobilis* variant of pyruvate decarboxylase can be used to create a shunt for pyruvate conversion to acetaldehyde. This exogenous expression was done in parallel with EM analysis and used to identify gene knockout targets that increased the yield of microbial ethanol production. Disruption of competing pyruvate pathways directed carbon flux towards the desired ethanol fermentation pathway. As seen with the carbon flux in Trinh et al. 2008, Shen et al. 2013 achieved a similar result with maximizing the yield of fermentation products by not only directing carbon flux but also by creating a NADH driving force by coupling the desired fermentation pathway with growth. The surplus of NADH created by disrupting competiting pathways was pulled through the desired pathway for increased productivity.

Kaczowka et al. 2005 provided evidence that the ZmPDC could replicate successfully in *H. volcanii*. This study used circular dichroism to show that ZmPDC expressed in *H. volcanii* was structurally similar to *E. coli* as displayed in Figure IIB-C at pH 6.5; recall that ZmPDC can produce ethanol in *E. coli* under anaerobic conditions (Trinh et al., 2008). ZmPDC gene products not only can fold correctly when expressed in *H. volcanii* but also show activity (in *E. coli*) in saline environments up to 4 M salt (Figure IIA). However, Kaczowka et al. 2005 observed acetate and lactate organic acids with no detectable amount of alcohol dehydrogenase activity in cell lysates suggesting that ethanol production is limited not only by pyruvate decarboxylase activity but alcohol dehydrogenase activity as well.

Moran-Reyna et al. 2014 investigated the differential expression of *H. volcanii* in extremes of pH (6.0 and 8.0). This study showed viability and growth in these varying pH ranges along with insight in the cellular

mechanisms that these environmental perturbations induce on *H. volcanii*. In the supplementary datasets of Moran-Reyna et al. 2014, alcohol dehydrogenase II (PFAMS: 00107, 08240; fold change: +3.33) was upregulated in the acidic conditions (pH 6.0). As mentioned above, Kaczowka et al. 2005 assayed ZmPDC structure in acidic conditions (pH 6.5) while growth assays were conducted in neutral pH environments.

Timpson et al. 2013 provided insight into the characteristics of alcohol dehydrogenase paralogs in *H. volcanii*. The study found that *adh2* (HvADH2) was highly robust in terms of temperature, ionic conditions, and stability. In agreement with Moran-Reyna et al. 2014, they also report that HvADH2 activity takes place in acidic conditions. The substrate specificity of HvADH2 is broad and has a higher affinity for acetaldehyde than HvADH1. The high affinity for acetaldehyde and NADH would aid in using a driving force-type approach by a higher probability of pulling the flux between ZmPDC to HvADH2 and ethanol production instead of the reverse reaction. As depicted in Table III, HvADH1 has a higher affinity for ethanol than HvADH2 so by using HvADH2 as the driving force, the likelihood for the reaction direction of acetaldehyde to ethanol is increased.

As Kaczowka et al. 2005 was on the right path in using ZmPDC with *H. volcanii* for engineering fermentation, together, these independent studies seem to contain the missing pieces of a complete picture:

3.1.1. Recall Kaczowka et al. 2005 conducted their *H. volcanii* experiments at pH 6.9 – pH 7.0 but their ZmPDC structural assays were measured in acidic pH 6.5 conditions. While evidence exists for ZmPDC expression in *H. volcanii,* there were little to no signs of enzymatic activity or ethanol production. They failed to measure any alcohol dehydrogenase activity at neutral pH but Moran-Reyna et al. 2014 and Timpson et al. 2013 show upregulation and enzymatic activity in acidic conditions, respectively. This connection introduces the possibility that the desired ethanol pathway was disrupted by insufficient pH environmental conditions (i.e. neutral instead of acidic).

3.1.2. Kaczowka et al. 2005 conducted previous work that indicates differences in codon usage between exogenous pyruvate decarboxylase genes and host organisms can diminish enzymatic funcationality (Talarico, Ingram, & Maupin-Furlow, 2001). The ZmPDC gene extensively uses GCU, AAU, CAU, CCU, UAU and GUU codons while these are less frequently used by the *H. volcanii*'s translational machinery. Previous studies have shown codon usage frequencies influence structural stability and proper protein folding Akashi, 1994; Quax, Claassens, Söll, & van der Oost, 2015; Yu et al., 2015). It is possible that low levels of ZmPDC and lack of activity resulted from incompatible usage between donor and host strains (*Z. mobilis* and *H. volcanii*, respetively).

3.1.3. Fermentation typically occurs in anaerobic conditions, Kaczowka et al. 2005 used aerobic (microaerobic at best) conditions during their fermentation experiments. Adjusting the oxygen availability in parallel with methods involving directed carbon flux and coupling a NADH-dependent driving force with ethanol production for growth makes it possible to obtain an ethanologenic strain of recombinant *H. volcanii.*

3.1.4. Kaczowka et al. 2005 reported acetate and lactate products indicating non-directed carbon flux towards competing fermentation pathways decreasing the resources for ZmPDC.

## 3.2. Hypotheses

**3.2.1**. Optimizing codon usage frequencies for ZmPDC by using the codon usage of both donor and host organisms would create a correctly folded stable enzymatically active ZmPDC enzyme in *H. volcanii*.

**3.2.2.** High-yield ethanol production can be achieved in a recombinant *H. volcanii* by virtue of: (i) expression of *H. volcanii*-optimized ZmPDC to create a shunt from pyruvate to acetaldehyde; (ii) direct carbon flux towards acetaldehyde and ethanol production by knocking out competing pathways; and (iii) creating a NADH-dependent driving force with HvADH2 and simultaneously coupling NAD+ regeneration to growth.

**3.2.3.** Adjusting environmental conditions to an acidic anaerobic environment, in conjunction with the above, would activate HvADH2 and produce ethanol.

## 3.3. Project Aims

The following project aims are designed to provide a resource for researchers using fluxomic and pathway analyses for the metabolic engineering of custom microbial strains.

### 3.3.1. Aim I | Identify gene targets to increase ethanol yield in *H. volcanii*

For my first aim, I will: (i) provide a manually curated conversion of all metabolites, genes, enzymes, and pathways from Trinh et al. 2008's EM analysis using reproducible KEGG identifiers showing all required pathways in their minimal ethanologenic strain; (ii) customize a System's Biology Markup Language [SBML] file containing the minimal set of metabolic reactions required for ethanol production using descriptive KEGG identifiers and annotations; (iii) identify *H. volcanii* orthologs of *E. coli* knockout targets identified by EM analysis for directed carbon flux; and (iv) use hidden markov models [HMMs] to identify additional *H. volcanii* genes containing enzymatic domains in the reactions specified by Trinh et al. 2008.

### 3.3.2. Aim II | Create software for optimal reverse-translation, database access, and format conversion

For my second aim, I will: (i) construct a command-line software tool in Python 2.7.11 for reverse-translation of donor amino-acid sequence using donor and host codon usage frequencies with options for motif exclusion and metrics containing GC-content, K-mer content, Shannon Entropy, isoelectric index, and Kyte-Doolittle hydrophobicity; (ii) provide a means for accessing the KEGG database within Python; and (iii) create tools for generating custom SBML and Genbank files.

### 3.3.3. Aim III | Synthesize components of knockout constructs for future iGEM members

For my third aim, I will: (i) construct KIKO2, a physical non-replicative plasmid containing a unique set of linkers for insertion; (ii) synthesize physical knockout inserts with disruption cassettes for lactate dehydrogenase and malic enzyme containing linkers for KIKO2; (iii) generate a file containing ZmPDC DNA sequences for optimal expression/translation in *H. volcanii* (with features

listed in 3.3.2.ii; and (iv) construct Genbank files for pTA963, KIKO2 with disruption cassettes for lactate dehydrogenase and malic enzyme to increase pyruvate and NADH availability.

### 3.4. Research Project

The ultimate goal of this research project is to synthesize a strain of ethanologenic *H. volcanii* by: (i) creating a direct metabolic route from pyruvate to acetaldehyde via pyruvate decarboxylase introduction; (ii) direct carbon flux towards acetaldehyde and ethanol production; and (iii) create a NADH-dependent driving force by coupling the desired fermentation pathway to growth. This research illuminates the engineering of a minimal ethanologenic strain of *H. volcanii* by presenting novel reverse-translation software and creating preliminary knockout constructs to increase the availability of pyruvate and NADH recycling. Engineering pyruvate decarboxylase production in an organism as robust as *H. volcanii*, in terms of salinity, pH, and temperature compatibility, is expected to progress the biofuel sector of biotechnology.

## 4.0 Methods | Results

Software, SBML, and Genbank files are hosted at the GitHub repository while pathway and ortholog tables are hosted at a stable Google Drive repository. All files are open-access; contact jol.espinoz@gmail.com for hosting issues.

### 4.1. Reverse-translation codon optimization software

Reverse-translation software was written entirely in *Python 2.7.11* using built-in modules and *BioPython* for rapid fasta sequence parsing. Genetic code conversion tables and versions reflected the notation of the NCBI 4.0 edition (ftp.ncbi.nih.gov/entrez/misc/data/gc.prt). The code can be accessed via private GitHub archive https://gist.github.com/jolespin/bc64edb0ef9aa488f13cb1be7b2bfb48. Reverse-translation products for 1000 variants available at https://docs.google.com/spreadsheets/d/1gptd6wQ7bi6eDLqzyYE3Y258Z3H0ZnxpC9rOBh-2qm4/.

### 4.2. Pathway and gene target selection

The knockout gene targets in *H. volcanii* were L-lactate dehydrogenase (Uniprot: D4GZF9) and Bifunctional malic enzyme oxidoreductase/phosphotransacetylase (Uniprot: D4GSQ3) for directed carbon flux and increased NADH availability. The metabolic pathway targeted by disrupting lactate dehydrogenase and malic enzyme were KEGG:hvo00620 (FEM3 and ANA2 in Trinh et al. 2008, respectively).

### 4.3 Conversion of pathways, metabolites, and genes

Database conversion tools were created using *Python 2.7.11* and *Orange-Bioinformatics* module (https://pypi.python.org/pypi/Orange-Bioinformatics/2.5.34).

Pathway conversion table based on KEGG identifiers from 3.3.1.i can be accessed at:

https://docs.google.com/spreadsheets/d/1YROR0Fd5s5-fiuoGpFeTgVW11a4j6eJ1Aex7PHJ8dKw/

System's Biology Markup Language metabolic system file from 3.3.1.ii available in Supplementary.

## 4.4. Hidden markov models for ortholog discovery

Ortholog-table as described in 3.3.1.iii. with gene cutoff scores and e-values for targets is available at:

https://docs.google.com/spreadsheets/d/1_JTuBKS1dMGX6BWJPi1oRlVNA4--W8Vb9Hr6DGGsEHs/

*Full_trusted* boolean field indicates if the score is above the trusted cutoff value determined by curated sources. *Full_noise* boolean field indicates if the score is above the noise cutoff value used for less confident but still probable candidate genes.

### 4.4.1. Identifying protein domains

Protein domains [i.e. HMMs including *TIGRFAMs* and *PFAMs*] for the knockout genes in *E. coli*, identified by Trinh et. al 2008, were extracted from the Uniprot database (http://uniprot.org/).

### 4.4.2. Retrieving seed sequences

HMM seed sequences were retrieved from *J. Craig Venter Institute's TIGRFAM* database (http://jcvi.org/cgi-bin/tigrfams/SeedAlignment.cgi?acc=).

### 4.4.3. Creating multiple sequence alignment

HMM seed sequences were aligned using *clustalw2*'s command-line tool with default settings (http://www.clustal.org/clustal2/).

### 4.4.4. Building and preparing hidden markov models

HMMs were constructed using *HMMER3's hmmbuild* and *hmmpress* tools (http://hmmer.org/) (Mistry, Finn, Eddy, Bateman, & Punta, 2013)

### 4.4.5. Constructing putative protein database

Putative protein database of open reading frames were extracted from *H. volcanii* chromosomes [NC_013964.1, NC_013965.1, NC_013966.1, NC_013967.1, and NC_013968.1] with *Glimmer2's* (Delcher, Harmon, Kasif, White, & Salzberg, 1999) *long-orfs* and *extract* tools using *atg,gtg,ttg* alternative start codons).

### 4.4.6. Searching putative proteins for target domains

Putative protein databases were scanned for target domains (i.e. domains from 4.3.1.) using *HMMER3's hmmscan* tool (Mistry et al., 2013).

## 4.5. Exogenous pyruvate decarboxylase overexpression components

The sequences generated from the above software were used to synthesize physical DNA inserts via gBlocks® Gene Fragments technology. Reverse-translation of pyruvate decarboxylase (Uniprot: P06672) sequence was employed using *Zymomonas mobilis subsp. mobilis (strain ATCC 31821/ZM4/CP4)* as the donor organism. The goal was to insert the fragment into the *H. volcanii* conditional overexpression vector pTA963 utilizing a tryptophan-inducible promoter *p.tnaA* (Allers, 2010) (Supp. Figure I, Genbank: https://gist.github.com/jolespin/2cfd75d311d683c6dd97679054d92b14) with F:*atgcaccaccaccaccaccac ATGTCGTACACGGTCGGC* and R: *tatcaagcttatcgacatTCAGAGGAGCTTG TTGACC* primers; lowercase indicates plasmid complementarity and uppercase binds the insert (Supp. Sequence I).

## 4.6. Knockout plasmid construction

In order to knockout genes in *H. volcanii*, a knockout template plasmid must be constructed. I created a version of pTA963 that lacked the *H. volcanii* replication of origin with the minimal amount of regions required for the disruption process; referred to as KIKO2 shown in Figure VIII (Genbank: https://gist.github.com/jolespin/fd271a4c2048aa69be37ea5c82743522). The primers used to amplify KIKO2 from pTA963 were F: *gaacgtggcgagaaaggaagggTGAGCAAAAGGCCAGCAA* and R: *cgccggctttccccgtcGCCCTTTGACGTTGGAGT* using *New England Biosciences* Q5 High-Fidelity DNA Polymerase at an annealing temperature of 66˚C according the manufacturer's protocol.

## 4.7. Knockout insert components

The knockout inserts for malic enzyme and lactate dehydrogenase (Supp. Sequences II and III, respectively) were synthesized using *gBlocks® Gene Fragments* technology with *GGCTAGTTTGGCGGATCTTAGCA GATACTACCCAGGTGCTTTCAACTACTTTGCTTGACGATTTACGCGCGCTATCCCGTAATCTTCAAATTA AAACATAGC* inserted in the middle of the gene to disrupt gene transcription. The primers used to amplify both malic enzyme and lactate dehydrogenase gene blocks were F: *GACGGGGAAAGCCGGCGC* and R: CCCTTCCTTTCTCGCCACGTTC with an annealing temperature of 66˚C using the same polymerase as the former PCR reaction. The plan for fusing the inserts with the knockout plasmid backbone was to use *Synthetic Genomics* Gibson Recombination kit to circularize the 2 linear fragments. After circularization, competent *E. coli* would be transformed using either electroporation or heat-shock and grown on Ampicillin-infused agar plates. Once colonies would be identified, colony PCR would be performed using *Clontech* Titanium Taq Polymerase to ensure proper integration. Candidate vectors would be isolated and sequenced.

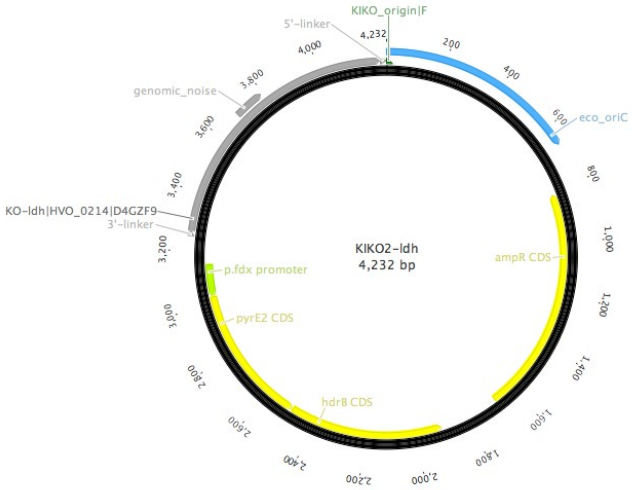


**Figure VIII** | KIKO2 Knockout Plasmid with D-Lactate dehydrogenase disruption cassette integration
Final labeled knockout plasmid product after fusing building block fragments
Genbank: https://gist.github.com/jolespin/1ffb6d652509507a972d571bff54e10d

## 5.0. Discussion

This research uses a combination of computational techniques to identify a cohort of gene knockout targets to direct carbon flux and to create a NADH-dependent driving force to couple ethanol production with growth. Elementary mode analysis was used to identify known genes and pathways in *H. volcanii* and *E. coli* (3.3.1.i-iii) that can be disrupted to limit the possible routes for pyruvate traversal, thus, directing carbon-rich substrate towards the desired fermentation destination. This same technique was employed to identify targets that will increase the pool of NADH redox molecules for maximizing HvADH2 productivity; essentially, a driving force pulling the flux into the ethanol pathway.

For further measures, this research employs hidden markov models to increase the gene target pool by identifying protein domains associated with pathway targets from elementary mode analysis in both annotated and theoretical *H. volcanii* proteins. This approach will provide a great resource for identifying future knockouts with the goal of either minimizing carbon sinks, eliminating non-essential NADH consumption, or compressing the genome size into a set of minimal genetic components. This method is robust in the idea that it is not limited to annotation availability or the possibility of incorrectly annotated proteins that could assume a different role.

The ortholog and metabolic reaction tables in conjunction with tools provided for creating custom SBML files (a widely used system's biology format) and accessing KEGG databases in a programming environment will greatly enhance the productivity for future researchers in metabolic engineering and synthetic biology-oriented foci. A persistent inconvenience in metabolic modeling, and virtually all of bioinformatics, is the inconsistency of identifiers and formats across the discipline. This research alleviates the conflict of manually searching alternative names, common names, paralogs, or chirality for genes, proteins, and compounds by using the comprehensive KEGG database and providing a tool for

custom applications.

Introducing exogenous genes is inherent in molecular and synthetic biology. It was once thought that high-fidelity translation was a direct result of highly abundant tRNAs and frequent codon usage (Akashi, 1994). However, recent studies have shown, with empirical evidence as support, that codon usage regulates protein function by affecting co-translational protein folding (Yu et al., 2015). In addition, computational approaches have revealed significant correlations between protein structural motifs and codon usage biases (Pechmann, Chartron, & Frydman, 2014). Recently, Tessa et al. 2016 discusses known types of codon bias and their influence on translation and protein elongation (Quax et al., 2015). Considering the increasing need for recombinant protein expression, the reverse-translation software accompanying this project provides an invaluable resource for generating a collection of sequences designed for optimal expression in a particular host. To address the issue of incorporating less frequent codons to stall the ribosome for proper folding, this software takes the original gene sequence from a donor organism and applies the codon rank towards reverse-translation using host species codon usage bias for an optimized sequence. For increased applicability, the software also provides an option for excluding codons that would introduce a particular set of motifs; an essential asset for virtually all molecular biology experiments. This software also provides metrics for evaluating sequence products such as GC-content, K-mer content, and Shannon entropy.

To advance this project further, empirical testing must evaluate the method's findings. This research provides individual components including inserts and plasmid backbones, with the appropriate linkers, that are synthesized, quantified, and experimentally ready. With proper funding and resources, the knockout plasmids could be constructed in less than a week's time with the high-fidelity DNA components generated by this project. Once the above are complete, growth assays in acidic anaerobic environments would be used to compare ethanol yield in a combinatorial approach. This can be extended by knocking out additional pyruvate sinks and, eventually, all redundant pathways non-essential for fermentation and rapid growth using transposon mutagenesis and fluxomics analysis. After empirical testing is complete, the optimal combination of knockouts and exogenous gene expression would be used to generate a stable minimal cell-line void of additional transformations and expression plasmids for standardized industrial use.

# 6.0. Future

## 6.1. Future of metabolic analysis

As mentioned above, elementary mode analysis is computationally expensive and difficult, if not impossible, to accomplish on standard 4 - 8GB memory laptop computers. Recent developments have been made that are expected to increase the availability of this technology to more researchers by increasing algorithm efficiency and expanding the framework of the algorithms.

For instance, this year researchers at *Universitat Politecnica de Valencia* in Spain have applied principal component analysis [PCA] to the growing field of fluxomics in an effort to decrease the dimensionality and complexity of metabolic analysis (Folch-Fortuny, Marques, Isidro, Oliveira, & Ferrer, 2016). Their main goals for using a modified PCA in fluxomics was to: (i) identify components of metabolism that describe the most variability; and (ii) relate the findings to the behavior of biological systems in terms of substrate consumption and protein production. The algorithm they developed, Prinicipal Elementary Mode Analysis [PEMA], finds the common set of active EMs in various flux distributions, thereby substantially reducing

the number of pathways to explain a complex fluxome.

The Planes group confront the computational challenge by devising a linear programming-based tree search algorithm, TreeEFM, that efficiently enumerates a subset of EMs in genome-scale metabolic networks (Pey et al., 2015). While most EM analysis studies focus on medium scale metabolomes, Pey et al. 2015 strives to increase versatility for incorporating this technology into both genome-scale and metagenomics-scale metabolic networks, where the number of chemical reaction exponentially increases. TreeEFM uses a linear program-based search procedure that is similar to the branch-and-bound approaches to solve integer programming models (Land & Doig, 2010). To increase the efficiency of calculating EMs, the group employs the following heuristic rules: (i) avoid recalculating previously identified EMs; and (ii) avoid solving nodes for linear programs without a solution. The weight of the computational time is spent on solving linear programs and this decreases the total number of computations required to obtain EMs.

Researchers from *Tufts University* in Massachusetts devised a graph-based network approach towards computing high dimensional metabolic network data (Ullah, Aeron, & Hassoun, 2016). Ullah et al. 2016 presents an algorithm, gEFM, that uses graph traversal to compute EMs. gEFM uses the Mavrovouniotis approach to transforms the initial set of reactions and iteratively incorporate the set of stoichiometric constraints associated with each metabolite into a complete set of pathways; satisfying all EM constraints.

As technology develops, compute resources become more attainable and the algorithms that use them increase in efficiency. Efforts to improve the obstacles that are inherently present in metabolic analysis and fluxomics will advance the understanding of kinetic-modeling and visualizing high-dimensional datasets. Using these methods in parallel with the transposon mutagenesis approach employed in the creation of minimal genomes (refer to 6.3. for details) towards the realm of biofuels and synthetic biology will progress our understanding of living systems and our ability to address some of the most pressing issues humanity faces.

## 6.2. Future of biofuels

A stable cell-line has hitherto been synthesized that is capable of toggling between growth and biofuel production. Filling this gap would make renewable energy more competitive; ultimately, reducing the rate of human-induced climate change. Using a fusion of informatics and genetic engineering techniques, a minimal strain of ethanologenic *H. volcanii* can be synthesized that reproduces only in the presence of oxygen while producing ethanol solely in anoxic conditions. Trinh et al. 2008 uses EM analysis to identify which pathways were relevant to growth and ethanol production in anaerobic environments. However, they did not create the schematics for a strain that could toggle between growth and ethanol production; a feature in high demand for standardizing industrial-scale procedures. This technology is designed for use as a renewable source of energy but would be relevant in medical settings and, of course, commercial alcohol production.

Further engineering steps could be employed to extend the organism to degrade and utilize the lignin component of lignocellulosic biomass. These genes could be identified from existing haloarchaeal annotations or discovered using hidden markov models in environmental metagenomic putative protein databases. Once a halophilic version of the lignin degradation system is revealed, the genes can be integrated into the ethanologenic strain produced from the above research for a fully autonomous ethanol producer.

While ethanol has several applications, butanol is the next step of biofuels largely due to its high 29.2 MJ/L energy density (Cheng, 2009). However, butanol is typically more toxic to cells than ethanol and requires more carbon substrate to form the product which makes it a more advanced product goal. Several groups have already addressed the challenge of generating butanol using metabolic engineering approaches (Atsumi et al., 2008; Nakayama et al., 2011; Shen et al., 2011).

The need for renewable energy and resources is vital for the growth of humanity. Biofuels are an excellent successor to gasoline because they can be used in many unmodified gasoline engines. There has yet to be a microbe optimized solely for the binary purpose of growth and fuel production. The current methods have not partitioned growth from fuel production and have not systematically quantified each reaction, compound, enzyme, and gene for standardized reproducibility. The progression of this research will optimize the efficiency of biofuel production utilizing resources ranging from excess plant matter in the agriculture industry to fast growing bamboo and rampant invasive species. The direct application of this research is to benefit our energy demanding civilization by making biofuels an exemplary fuel that sets precedence for future sustainable energy sources and, ultimately, the rescinding of industrial-mediated global warming.

### 6.3. Future of synthetic biology

Synthetic biology is a rapidly growing field of interdisciplinary science that serves as a junction between biochemistry, computational science, molecular biology, and philosophy. Synthesizing novel strains of an organism for biotechnology is a clear application for both monetary gain and advancing science. However, groups lead by J. Craig Venter and Neri Oxman seek to project this research into a higher dimensional framework. For example, this year researchers at the *J. Craig Venter Institute* have synthesized a minimum microbial organism (*JCVI-syn3.0*) based on a *Mycoplasma mycoides* genome (Hutchison et al., 2016). The researchers use transposon mutagenesis to reveal classes of quasi-essential and essential genes that are required for robust growth. The *JCVI-syn3.0* organism contains a genome of 531 kb and 473 genes; substantially smaller than any autonomously replicating cell found in nature. Interestingly, of these 473 genes, 149 of which were not yet annotated and had unknown biological function. This approach to synthetic biology can give insight into mutation, adaptation, and unknown, yet essential, pathways required for all life to exist. As my research continues, I plan to extend this concept to minimal ecological systems and further minimize *JCVI-syn3.0* by optimizing sequence overlap to physically compress the genome. In addition to evolutionary studies arguably "playing God", Neri Oxman and colleagues use synthetic biology in combination with microfluidics and 3D-printing to produce inexpensive materials used for engineering, fashion, and green architecture (Patrick et al., 2015). For exploratory applications, synthetic biology has even been proposed as a method for reengineering humans to withstand extended space exploration missions (Nip, 2015).

The inherent theme of these types of approaches contribute to an in-depth understanding of the basis of microbial metabolism, what the basis of cellular life is, and how a microbe can be modified to optimize the production of a selected metabolite. As eloquently asked by Dr. Brett Olivier in his *PySCeS* thesis, "[w]ould Leonardo, if alive, be building digital cells? As technology has developed, so levers and pulleys have given way to ones and zeros; mechanization is now encapsulated by digitization". Understanding the basis of life at the system's level would provide the grounds for engaging intimately into the art of designing organisms for specific roles in helping humanity not only fuel our economy with renewable energy, but also to restore the ecosystem and defend against pathogens via novel therapeutics.

## 6.0. Supplementary Material

Refer to the accompanying supplementary document for files and *GitHub* archives.

## 7.0. Acknowledgments

I would like to thank: David Bernick for the opportunity to apply new computational approaches while guiding me through the synthetic biology framework; Chad Salktikov for the great lectures and insight on microbes and redox reactions; Josh Stuart for introducing the computational methods used in system's biology; Kevin Karplus for leveling up my programming abilities; Christopher Dupont and Alexander Richter from *JCVI* for the idea of creating ortholog tables with hidden markov models; Miles F. Wilkinson, Chih-Hong Lou, and Eleen Y. Shum from *UCSD* for their patience and scientific wisdom; my mother Robbin, my girlfriend Madison, and my friends for their support; and lastly, the Nature of Big Sur's expanse, Highway 1's ocean cliffs, and Santa Cruz's coastal *Sequoia sempervirens* forests that continue to inspire my quest in understanding the universe around me.

## 8.0. References

Acuña, V., Chierichetti, F., Lacroix, V., Marchetti-Spaccamela, A., Sagot, M.-F., & Stougie, L. (2009). Modes and cuts in metabolic networks: complexity and algorithms. *Bio Systems*, *95*(1), 51–60. http://doi.org/10.1016/j.biosystems.2008.06.015

Akashi, H. (1994). Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy. *Genetics*, *136*(3), 927–935.

Allers, T. (2010). Overexpression and purification of halophilic proteins in Haloferax volcanii. *Bioengineered Bugs*, *1*(4), 288–290. http://doi.org/10.4161/bbug.1.4.11794

Atsumi, S., Cann, A. F., Connor, M. R., Shen, C. R., Smith, K. M., Brynildsen, M. P., … Liao, J. C. (2008). Metabolic engineering of Escherichia coli for 1-butanol production. *Metabolic Engineering*, *10*(6), 305–311. http://doi.org/10.1016/j.ymben.2007.08.003

Baliga, N. S., Bjork, S. J., Bonneau, R., Pan, M., Iloanusi, C., Kottemann, M. C. H., … DiRuggiero, J. (2004). Systems level insights into the stress response to UV radiation in the halophilic archaeon Halobacterium NRC-1. *Genome Research*, *14*(6), 1025–1035. http://doi.org/10.1101/gr.1993504

Brown, M. E., & Chang, M. C. (2014). Exploring bacterial lignin degradation. *Current Opinion in Chemical Biology*, *19*, 1–7. http://doi.org/10.1016/j.cbpa.2013.11.015

Carroll, A., & Somerville, C. (2009). Cellulosic Biofuels. *Annual Review of Plant Biology*, *60*(1), 165–182. http://doi.org/10.1146/annurev.arplant.043008.092125

Center for Sustainable Systems, (University of Michigan). (2015, October). U.S. Renewable Energy Factsheet. University of Michigan. Retrieved from http://css.snre.umich.edu/css_doc/CSS03-12.pdf

Cheng, J. (2009, December 1). Biomass to Renewable Energy Processes. Retrieved May 7, 2016, from https://www.crcpress.com/Biomass-to-Renewable-Energy-Processes/Cheng-Cheng/9781420095173

Coker, J. A., DasSarma, P., Kumar, J., Müller, J. A., & DasSarma, S. (2007). Transcriptional profiling of the model Archaeon Halobacterium sp. NRC-1: responses to changes in salinity and temperature. *Saline Systems*, *3*, 6. http://doi.org/10.1186/1746-1448-3-6

Coquelle, N., Talon, R., Juers, D. H., Girard, E., Kahn, R., & Madern, D. (2010). Gradual adaptive changes of a protein facing high salt concentrations. *Journal of Molecular Biology*, *404*(3), 493–505. http://doi.org/10.1016/j.jmb.2010.09.055

Delcher, A. L., Harmon, D., Kasif, S., White, O., & Salzberg, S. L. (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Research*, *27*(23), 4636–4641.

DeVeaux, L. C., Müller, J. A., Smith, J., Petrisko, J., Wells, D. P., & DasSarma, S. (2007). Extremely radiation-resistant mutants of a halophilic archaeon with increased single-stranded DNA-binding protein (RPA) gene expression. *Radiation Research*, *168*(4), 507–514. http://doi.org/10.1667/RR0935.1

Dumsday, G. J., Zhou, B., Yaqin, W., Stanley, G. A., & Pamment, N. B. (1999). Comparative stability of ethanol production by Escherichia coli KO11 in batch and chemostat culture. *Journal of Industrial Microbiology & Biotechnology*, *23*(1), 701–708.

Eschbach, M., Schreiber, K., Trunk, K., Buer, J., Jahn, D., & Schobert, M. (2004). Long-Term Anaerobic Survival of the Opportunistic Pathogen Pseudomonas aeruginosa via Pyruvate Fermentation. *Journal of Bacteriology*, *186*(14), 4596–4604. http://doi.org/10.1128/JB.186.14.4596-4604.2004

Flam, F. (1994). The chemistry of life at the margins. *Science*, *265*(5171), 471–472. http://doi.org/10.1126/science.8036489

Folch-Fortuny, A., Marques, R., Isidro, I. A., Oliveira, R., & Ferrer, A. (2016). Principal elementary mode analysis (PEMA). *Molecular bioSystems*, *12*(3), 737–746. http://doi.org/10.1039/c5mb00828j

Gladden, J. M., Park, J. I., Bergmann, J., Reyes-Ortiz, V., D'haeseleer, P., Quirino, B. F., … Singer, S. W. (2014). Discovery and characterization of ionic liquid-tolerant thermophilic cellulases from a switchgrass-adapted microbial community. *Biotechnology for Biofuels*, *7*(1), 15. http://doi.org/10.1186/1754-6834-7-15

Hutchison, C. A., Chuang, R.-Y., Noskov, V. N., Assad-Garcia, N., Deerinck, T. J., Ellisman, M. H., … Venter, J. C. (2016). Design and synthesis of a minimal bacterial genome. *Science (New York, N.Y.)*, *351*(6280), aad6253. http://doi.org/10.1126/science.aad6253

Iding, H., Siegert, P., Mesch, K., & Pohl, M. (1998). Application of alpha-keto acid decarboxylases in biotransformations. *Biochimica Et Biophysica Acta*, *1385*(2), 307–322.

Ingram, L. O., & Conway, T. (1988). Expression of Different Levels of Ethanologenic Enzymes from Zymomonas mobilis in Recombinant Strains of Escherichia coli. *Applied and Environmental Microbiology*, *54*(2), 397–404.

Jungreuthmayer, C., & Zanghellini, J. (2012). Designing optimal cell factories: integer programming couples elementary mode analysis with regulation. *BMC Systems Biology*, *6*, 103. http://doi.org/10.1186/1752-0509-6-103

Kaczowka, S. J., Reuter, C. J., Talarico, L. A., & Maupin-Furlow, J. A. (2005). Recombinant production of Zymomonas mobilis pyruvate decarboxylase in the haloarchaeon Haloferax volcanii. *Archaea (Vancouver, B.C.), Archaea*, *1*, *1*(5, 5), 327, 327–334. http://doi.org/10.1155/2005/325738

Kennedy, S. P., Ng, W. V., Salzberg, S. L., Hood, L., & DasSarma, S. (2001). Understanding the adaptation of Halobacterium species NRC-1 to its extreme environment through computational analysis of its genome sequence. *Genome Research*, *11*(10), 1641–1650. http://doi.org/10.1101/gr.190201

Kim, J.-H., Block, D. E., & Mills, D. A. (2010). Simultaneous consumption of pentose and hexose sugars: an optimal microbial phenotype for efficient fermentation of lignocellulosic biomass. *Applied Microbiology and Biotechnology*, *88*(5), 1077–1085. http://doi.org/10.1007/s00253-010-2839-1

Land, A. H., & Doig, A. G. (2010). An Automatic Method for Solving Discrete Programming Problems. In M. Jünger, T. M. Liebling, D. Naddef, G. L. Nemhauser, W. R. Pulleyblank, G. Reinelt, … L. A. Wolsey (Eds.), *50 Years of Integer Programming 1958-2008* (pp. 105–132). Springer Berlin Heidelberg. Retrieved from http://link.springer.com/chapter/10.1007/978-3-540-68279-0_5

Margesin, R., & Schinner, F. (2001). Potential of halotolerant and halophilic microorganisms for biotechnology. *Extremophiles: Life Under Extreme Conditions*, *5*(2), 73–83.

Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., & Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Research*, *41*(12), e121–e121. http://doi.org/10.1093/nar/gkt263

Moran-Reyna, A., & Coker, J. A. (2014). The effects of extremes of pH on the growth and transcriptomic profiles of three haloarchaea. *F1000Research*. http://doi.org/10.12688/f1000research.4789.2

Müller, J. A., & DasSarma, S. (2005). Genomic analysis of anaerobic respiration in the archaeon Halobacterium sp. strain NRC-1: dimethyl sulfoxide and trimethylamine N-oxide as terminal electron acceptors. *Journal of Bacteriology*, *187*(5), 1659–1667. http://doi.org/10.1128/JB.187.5.1659-1667.2005

Nakayama, S., Kiyoshi, K., Kadokura, T., & Nakazato, A. (2011). Butanol Production from Crystalline Cellulose by Cocultured Clostridium thermocellum and Clostridium saccharoperbutylacetonicum N1-4. *Applied and Environmental Microbiology*, *77*(18), 6470–6475. http://doi.org/10.1128/AEM.00706-11

Neale, A. D., Scopes, R. K., Wettenhall, R. E., & Hoogenraad, N. J. (1987). Pyruvate decarboxylase of Zymomonas mobilis: isolation, properties, and genetic expression in Escherichia coli. *Journal of Bacteriology*, *169*(3), 1024–1028.

Nip, L. (2015, November). *How humans could evolve to survive in space*. Retrieved from http://www.ted.com/talks/lisa_nip_how_humans_could_evolve_to_survive_in_space

Ollivier, B., Caumette, P., Garcia, J. L., & Mah, R. A. (1994). Anaerobic bacteria from hypersaline environments. *Microbiological Reviews*, *58*(1), 27–38.

Patrick, W. G., Nielsen, A. A. K., Keating, S. J., Levy, T. J., Wang, C.-W., Rivera, J. J., … Kong, D. S. (2015). DNA Assembly in 3D Printed Fluidics. *PloS One*, *10*(12), e0143636. http://doi.org/10.1371/journal.pone.0143636

Pechmann, S., Chartron, J. W., & Frydman, J. (2014). Local slowdown of translation by nonoptimal codons promotes nascent-chain recognition by SRP in vivo. *Nature Structural & Molecular Biology*, *21*(12), 1100–1105. http://doi.org/10.1038/nsmb.2919

Pey, J., Villar, J. A., Tobalina, L., Rezola, A., García, J. M., Beasley, J. E., & Planes, F. J. (2015). TreeEFM: calculating elementary flux modes using linear optimization in a tree-based algorithm. *Bioinformatics (Oxford, England)*, *31*(6), 897–904. http://doi.org/10.1093/bioinformatics/btu733

Quax, T. E. F., Claassens, N. J., Söll, D., & van der Oost, J. (2015). Codon Bias as a Means to Fine-Tune Gene Expression. *Molecular Cell*, *59*(2), 149–161. http://doi.org/10.1016/j.molcel.2015.05.035

Shen, C. R., Lan, E. I., Dekishima, Y., Baez, A., Cho, K. M., & Liao, J. C. (2011). Driving forces enable high-titer anaerobic 1-butanol synthesis in Escherichia coli. *Applied and Environmental Microbiology*, *77*(9), 2905–2915. http://doi.org/10.1128/AEM.03034-10

Talarico, L. A., Ingram, L. O., & Maupin-Furlow, J. A. (2001). Production of the Gram-positive Sarcina ventriculi pyruvate decarboxylase in Escherichia coli. *Microbiology (Reading, England)*, *147*(Pt 9), 2425–2435. http://doi.org/10.1099/00221287-147-9-2425

Timpson, L. M., Alsafadi, D., Mac Donnchadha, C., Liddell, S., Sharkey, M. A., & Paradisi, F. (2012). Characterization of alcohol dehydrogenase (ADH12) from Haloarcula marismortui, an extreme halophile from the Dead Sea. *Extremophiles: Life Under Extreme Conditions*, *16*(1), 57–66. http://doi.org/10.1007/s00792-011-0405-0

Timpson, L. M., Liliensiek, A.-K., Alsafadi, D., Cassidy, J., Sharkey, M. A., Liddell, S., … Paradisi, F. (2013). A comparison of two novel alcohol dehydrogenase enzymes (ADH1 and ADH2) from the extreme halophile Haloferax volcanii. *Applied Microbiology and Biotechnology*, *97*(1), 195–203. http://doi.org/10.1007/s00253-012-4074-4

Trinh, C. T., Unrean, P., & Srienc, F. (2008). Minimal Escherichia coli cell for the most efficient production of ethanol from hexoses and pentoses. *Applied and Environmental Microbiology*, *74*(12), 3634–3643. http://doi.org/10.1128/AEM.02708-07

Trinh, C. T., Wlaschin, A., & Srienc, F. (2009). Elementary Mode Analysis: A Useful Metabolic Pathway Analysis Tool for Characterizing Cellular Metabolism. *Applied Microbiology and Biotechnology*, *81*(5), 813–826. http://doi.org/10.1007/s00253-008-1770-1

Ullah, E., Aeron, S., & Hassoun, S. (2016). gEFM: An Algorithm for Computing Elementary Flux Modes Using Graph Traversal. *IEEE/ACM Transactions on Computational Biology and Bioinformatics / IEEE, ACM*, *13*(1), 122–134. http://doi.org/10.1109/TCBB.2015.2430344

Yu, C.-H., Dang, Y., Zhou, Z., Wu, C., Zhao, F., Sachs, M. S., & Liu, Y. (2015). Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. *Molecular Cell*, *59*(5), 744–754. http://doi.org/10.1016/j.molcel.2015.07.018

Zanghellini, J., Ruckerbauer, D. E., Hanscho, M., & Jungreuthmayer, C. (2013). Elementary flux modes in a nutshell: properties, calculation and applications. *Biotechnology Journal*, *8*(9), 1009–1016. http://doi.org/10.1002/biot.201200269