Modeling of a Cell-Free Synthetic System for Biohydrogen Production

Nicolas Fontaine, Brigitte Grondin-Perez, Frédéric Cadet, Bernard Offmann

Journal of
Computer Science & Systems Biology

Fontaine et al., J Comput Sci Syst Biol 2015, 8:3
DOI: 10.4172/jcsb.1000181

Research Article Open Access

Modeling of a Cell-Free Synthetic System for Biohydrogen Production

Nicolas Fontaine[1], Brigitte Grondin-Perez[2], Frederic Cadet[1,4] and Bernard Offmann[3,4]*

[1]University of La Reunion, Faculty of Sciences and Technology, INSERM, UMR S-1134, DSIMB, Laboratory of Excellence, LABEX GR, 97444 St Denis cedex, France
[2]University of La Reunion, Faculty of Sciences and Technology, Laboratoire d'Energétique, d'Electronique, et Procédés (LE2P), EA 4079, 97444 St Denis cedex, France
[3]University of Nantes, Faculty of Science and Techniques, Unité Fonctionnalité et Ingénierie des Protéines, 2, rue de la Houssinière, UMR-CNRS n°6286, BP 92208, 44322 Nantes cedex 3, France
[4]Peaccel Inc., 185 Alewife Brook Parkway #410, Cambridge, MA 02138, USA

Abstract

Hydrogen is a good candidate for the next generation fuel with a high energy density and an environment friendly behavior in the energy production phase. Micro-organism based biological production of hydrogen currently suffers low hydrogen production yields because the living cells must sustain different cellular activities other than the hydrogen production to survive. To circumvent this, teams have explored the synthetic assembly of enzymes in-vitro in cell-free systems with specific functions. Such a synthetic cell-free system was recently devised by combining 13 different enzymes to synthesize hydrogen from cellulose or cellobiose with better yield than microorganism-based systems. We used methods based on differential equations calculations to investigate how the initial conditions and the kinetic parameters of the enzymes influenced the productivity of a such system and, through simulations, identified those conditions that would optimize hydrogen production starting with cellobiose as substrate. Further, if the kinetic parameters of the component enzymes of such a system are not known, we showed how, using artificial neural network, it is possible to identify alternative models that account for the rate of production of hydrogen. This work demonstrates how modeling can help in designing and characterizing cell-free systems in synthetic biology. A web-based simulator implementing our differential equations based model is provided freely as a service for non-commercial usage at http://www.bo-protscience.fr/h2.

Keywords: Cell-Free synthetic system; Biochemical engineering; Hydrogen production; Mathematical modeling; Simulation; System optimization; Artificial neural network

Introduction

Hydrogen (H_2) is an alternative fuel in development, clean during combustion phase and with a high energy yield (142.35 kJ/g). Researchers study production of hydrogen from renewable biomass and different processes. Biological production is based on exploitation of micro-organisms with different biochemical pathways proficient at hydrogen production. Different pathways exist to produce hydrogen: dark fermentation from sugar, biophotolysis from absorption of light and water, and photo fermentation from absorption of light and organic acid [1-3]. Each pathway has a specific enzymes set but they suffer of a low yield hydrogen production. Indeed the microorganisms with such pathway are not designed to produce only hydrogen; other essential metabolic activities must be sustained for cell survival. A more recent biological approach, called cell-free synthetic pathway biotransformation [4], consists first to identify different target enzymes from different natural pathways, and next to combine them together in a synthetic system specific for one function such as hydrogen production. Different synthetic constructions of enzymes *in-vitro* in cell-free systems were build up for hydrogen production from the utilization of sugar as raw material [5-7]. In 2007, Zhang et al. have built a cell-free system of 13 enzymes in order to synthesize hydrogen from starch [7]. In 2009, Ye et al. have made a similar system of 13 enzymes but using as raw material, cellobiose the dominant product of enzymatic cellulose hydrolysis [5]. The hydrogen yield for such system is near 11.3 moles of H_2 per mole of glucose coming from the transformation cellobiose [5]. This yield represents 93% of the theoretical maximum of hydrogen production from glucose.

However, the rate of H_2 production of such a system is too weak in order to be competitive for an industrial scale. Some works have pointed out that this rate could potentially be enhanced 1000-fold if the enzyme kinetics could be improved [8]. Our study explored different modeling approaches to have a better understanding and control on hydrogen cell-free system with the objective to find different paths to improve the performance of production of hydrogen. First, a numerical model of the Ye et al. [5] system consuming cellobiose, based on ordinary differential equations, was built up in order to follow the system behavior in different conditions. The differential equations of the model have been written on the basis of biochemical knowledge for each enzyme reaction and their kinetic parameters. Then, through simulations, we searched for initial conditions and kinetic parameters that would optimize the production rate. Second, another modeling approach based on artificial neural network was developed to reproduce the global behavior of the system, in particular to predict the rate of production of hydrogen as a function of initial concentrations of substrates and enzymes. This approach doesn't require information on enzyme kinetics.

The overall aim of our investigation is to show that is it possible through modeling to find optimal conditions for the hydrogen production using the Ye et al. cell-free system [5] but also to show the suitability of modeling for the management of cell-free system in general.

***Corresponding author:** Bernard Offmann, University of Nantes, Faculty of Science and Techniques, Unité Fonctionnalité et Ingénierie des Protéines, 2, rue de la Houssinière, UMR-CNRS n°6286 , BP 92208, 44322 Nantes cedex 3, France, Tel: 33251125721; E-mail: bernard.offmann@univ-nantes.fr

Received February 21, 2015; **Accepted** March 28, 2015; **Published** March 30, 2015

Citation: Fontaine N, Grondin-Perez B, Cadet F, Offmann B (2015) Modeling of a Cell-Free Synthetic System for Biohydrogen Production. J Comput Sci Syst Biol 8: 132-139. doi:10.4172/jcsb.1000181

Method

Knowledge-based model construction

The basis of our work is the cell-free system that produces hydrogen from cellobiose consumption at 32°C, as developed by Ye et al. [5] (Figure 1). Our knowledge-based model was built from pre-existing models [5,9]. These models are based on ordinary differential equations (ODE) that characterize reaction mechanisms in the system. The system is composed of 19 metabolites, 13 enzymes that catalyze 14 reactions. Likewise, our knowledge-based model relied on the detailed knowledge of the kinetic features of each reaction and enzyme of the system. The 13 enzymes are the same as described in Ye et al. [5]: cellobiose phosphorylase (E1), phosphoglucomutase (E2), glucose-6-phosphate dehydrogenase (E3), 6-phosphogluconic dehydrogenase (E4), ribose 5-phosphate isomerase (E5), ribulose-5-phosphate 3-epimerase (E6), transketolase (E7), transaldolase (E8), triose-phosphate isomerase (E9), aldolase (E10), fructose-1,6-bisphosphatase (E11), phosphoglucose Isomerase (E12), $NADP^+$ dependent hydrogen dehydrogenase (E13). All 13 enzymes and the 14 associated reactions are further detailed in Supplementary Table 1.

Reactions can be represented under a numerical form using (i) the kinetic equations which are derived from the knowledge of the kinetic laws that govern the enzymes involved and (ii) the kinetic parameters associated with the enzymes: k_{cat}, the turnover number, the number of times each enzyme site converts substrate to product per unit time; K_m, the Michaelis-Menten constant, the affinity of the enzyme for substrate; K_i, the dissociation constant for inhibitor binding; K_{eq}, the equilibrium constant.

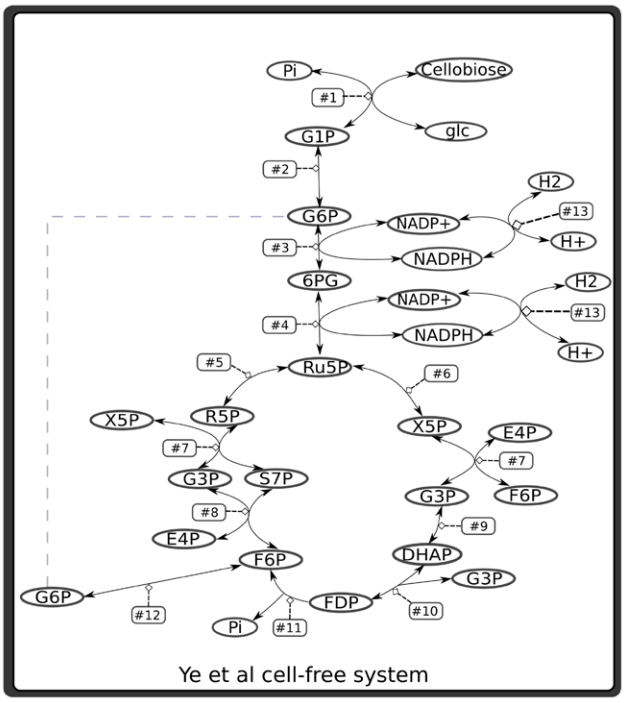

Figure 1: Detailed diagram of the cell-free system for biohydrogen production from cellobiose used by Ye et al. [5].

G1P, glucose-1-phosphate; G6P, glucose-6-phosphate; 6PG, 6-phosphogluconate; Ru5P, ribulose-5- phosphate; Pi, inorganic phosphate; R5P, ribose-5-phosphate; X5P, xylulose-5-phosphate; S7P, sedoheptulose-7-phosphate; E4P ´erythrose-4-phosphate; G3P, glyceraldehyde-3-phosphate; DHAP, dihydroxyacetone phosphate; F6P, fructose-6-phosphate; FDP, fructose-1,6-phosphate; #1,cellobiosephosphorylase; #2, phosphoglucomutase; #3, G-6-P dehydrogenase; #4, 6-phosphogluconate dehydrogenase ; #5 Phosphoriboseisomerase; #6, Ribulose 5-phosphate epimerase; #7, Transaldolase; #8, Transketolase; #9, Triose phosphate isomerase; #10, Aldolase, #11, Phosphoglucoseisomerase; #12, Fructose-1, 6-bisphosphatase; #13, NADP(+) hydrogen dehydrogenase.

Parameters screened for optimization	Final hydrogen concentration (mM)
No optimization	112
Enzyme degradation (λ)	118
Enzyme affinity for substrate (K_m)	238
Initial enzyme concentration (E_0)	591
Enzyme turn-over (kcat)	810

Table 1: Hydrogen final concentration value reached by the system after single stage optimization runs. E_0, k_{cat}, K_m and λ for all the 13 enzymes were optimized each separately while the other parameters were kept constant at their default values (see Method). The modifications were evaluated in order to obtain the best final hydrogen concentration (mM) after a simulation time of 9000 min with starting concentrations of 70 mM for cellobiose, 70 mM for inorganic phosphate and 1 mM for NADP. Additional results are featured in Figure 3.

Parameters screened for optimization	Selected enzymes	Final hydrogen concentration (mM)
Initial enzyme concentration (E_0)	E3, E4, E13	582
Enzyme turn-over (kcat)	E3, E4, E13	800
Enzyme affinity for substrate (K_m)	E4 and E13	236

Table 2: Hydrogen final concentration value reached by the system after single stage optimization runs. Here, E_0, k_{cat} and K_m for only E3 (glucose-6-phosphate dehydrogenase), E4 (6-phosphogluconic dehydrogenase) and E13 (NADP+ dependant hydrogen dehydrogenase) were optimized each separately while the parameters for the other enzymes were kept constant at their default values (see Method). The modifications were evaluated in order to obtain the best final hydrogen concentration (mM) after a simulation time of 9000 min with starting concentrations of 70 mM for cellobiose, 70 mM for inorganic phosphate and 1 mM for NADP. Additional results are featured in Figure 3.

First optimization (performed on all enzymes)	Second optimization	Optimal H_2 concentration (mM)
Initial enzyme concentration (E_0)	K_m (all enzymes)	827 (98.5%)
	λ (all enzymes)	815 (97.0 %)
	Km (only E4 and E13)	819 (97.5 %)
	λ (only E4 and E13)	793 (94.5 %)

Table 3: Hydrogen final concentration reached by the system after double stage optimization.

In the first stage optimization all the enzymes are modified on the initial concentration (E_0). The modifications are evaluated in order to obtain the best hydrogen final concentration (mM) from E0 modifications. Next, the optimized value for E_0 was integrated in the system for a second optimization round on another optimization parameter, K_m or λ. Simulations of 9000 min were executed with starting concentrations of 70mM for cellobiose, 70mM for phosphate and 1mM for NADP. Percentages between parentheses are with respect to maximum theoretical H_2 concentration that could be achieved.

Our kinetic equations integrate an enzyme degradation constant (λ) in order to reproduce the inactivation and the temporal degradation of enzymes. The kinetic equation and the kinetic parameters of each enzyme are described in the Supplementary Table 2. These equations were used to write the mass balance equations of each metabolite (Supplementary Table 3). Mass balance equations are ODE. The system has 19 metabolites, hence 9 ODE are required to follow all system metabolites. The solution of these 19 ODE permit to monitor the concentration of each metabolite as a function of time throughout simulation runs.

The model was conceptualized under the SBML format (Systems Biology Markup Language) that is a standard to represent biochemical networks [10]. Cell Designer and Copasi are software tools used for the model implementation in SBML format. The resolution of ODE was performed by the SBML ODE Solver Library or Copasi ODE solver.

System optimization

Optimization consists in modifying parameters of a system in order to improve its functioning and performance. At the end of an optimization process, the parameters of the system have optimized values. Our cell free system optimization was carried out to improve the final concentration of hydrogen. We selected four enzyme parameters for optimization: enzyme initial concentration (E_0), k_{cat}, K_m and degradation constant (λ).

We performed both single stage optimization and double stage optimization. Single stage optimization is carried out on only one parameter while the others are kept constant. Double stage optimization required two steps. The first step is a single stage optimization on one parameter. Next, the optimized value of this parameter was integrated in the system for a second optimization round on another parameter. In total, four single stage optimizations were accomplished, one for each optimization parameter (E_0, k_{cat}, K_m and λ). Two double stages optimizations called "E_0+K_m" and "E_0+λ" were performed: in "E_0+K_m", initial enzyme concentration (E_0) was optimized in first stage and substrate affinity (K_m) in second stage; in "E_0+λ", E_0 was optimized first and enzyme degradation constant (λ) in second stage. The modification of these parameters were modified within the following predefined limits: for E_0, between 1 and 100 U/mL; for k_{cat}, between 0.001 and 1 mmol.ml/min/U; for K_m, between Km_0/10 and 10Km_0 where Km_0 is the native experimentally derived value as featured in Supplementary Table 2; for λ, between 5 h and 75 h of half-life time. Otherwise, the default initial enzyme concentration (E0) was set to 10 U/mL and the default values for k_{cat}, K_m, λ were those indicated in Supplementary Table 2. All optimization simulations were started with 70 mM of cellobiose, 70 mM of phosphate, 1 mM of NADP. Each simulation run represented 9000 min (150 h) of the cell-free system operation. The optimization process was performed with the optimization module of Copasi using genetic algorithm as solver. Copasi have a set of default values for genetic algorithm that are adapted at the optimization of biochemical kinetics: 200 generations, 20 population size. Random number generator [11] takes care of the introduction of genetic variation by mutation and cross-over. This set-up allows an optimization process with acceptable solution in 5-10 min.

In theory, genetic algorithm could provide the best solution to a problem but requires a lot of computing time, an infinity period in extreme case. In respectable time delay, genetic algorithm doesn't give the perfect solution but it covers a set of solution near the optimal solution. For each operation of optimizations, multiple optimization simulations, at least 5, were carried.

Artificial neural network modeling

The model above used for optimization was built on prior knowledge of the kinetic laws and properties associated with each enzyme in the system and the definition of ordinary differential equations. However, sometimes this knowledge is incomplete. In this case, the construction of a robust ODE model is very difficult and it is at stake to find other modeling approaches. One such method is modeling using artificial neural networks (ANN). ANN models require only an initial dataset of input and output vectors. ANN models are very dependent of the number experiments available for the construction of the initial dataset. The size of dataset must be large enough to build an ANN model. If the size is too small, ANN method is not relevant.

After validation, ANN model allows to predict the value of output vectors from a set of input vectors. System predictions from ANN method are the most valid and relevant in the system conditions encountered in the initial dataset. Prediction from of out-of-the-box conditions must be subject to caution. The more the distance between the conditions of the dataset and the out-of-the-box conditions, the higher is the probability to have inaccurate and even false predictions.

Here we explored the use of ANN to model cell-free systems. Our aim is to validate its application for modeling output of cell-free systems knowing only the starting state of the system.

Artificial neural network is a computational model based on biological neural networks. The model structure is similar to the structure of the biological neural networks with the presence of computational units interconnected between them. One computational unit is assimilated to a formal neuron. A formal neuron can receive one or several input signals. The neuron treats the weighted summation of input signals to produce an output signal by the action of an activation function. Artificial neural network have a learning phase to identify a model with the adequate weights so that the model outputs match the system output. Because they can approximate nonlinear functions between inputs and outputs from incomplete databases and with specific accuracy, ANNs are now commonly used to model complex systems. The construction of an ANN model is achieved following four classical steps: (i) building of a database, (ii) determining model structure, (iii) fitting parameters and (iv) validating model. Each of these steps is detailed below.

In first instance, the knowledge-based model, as described in the previous section, was used to generate a database of hydrogen production trajectories that were simulated with different sets of starting conditions for the system. The starting conditions for the simulations, i.e the initial concentrations of the 13 enzymes and 3 substrates (NADP, cellobiose and inorganic phosphate), served as inputs for the ANN while the observed initial rate of hydrogen production at t=0 min and final hydrogen concentration reached at the end of the simulations served as outputs for the ANN. These inputs were chosen because they can be realistically set up experimentally. To constitute the database, initial values for the different simulation runs were varied between the following ranges: between 5 and 20 U/mL for enzyme concentration, between 0.5 and 2 mM for NADP concentration and between 35 and 140 mM for both cellobiose and inorganic phosphate concentrations.

The database comprised four distinct datasets. These are further detailed below. A first very large dataset, hereby called "Base zero", was constituted for the purpose of evaluating whether ANN can indeed be used and to build a model that can account for the rate of H_2 production and for the final level of H_2 in the system. The starting conditions that were used to build this dataset consisted in setting two possible values for each the 16 inputs and generating all possible combinations of these starting values, hence making a total of 2^{16}=65,536 starting conditions. Starting enzyme concentrations were set to 5 and 20 U/mL, NADP concentrations to 0.5 and 2 mM and for both cellobiose and phosphate to 35 and 140 mM. After validating the use of ANN to model both production rate and final concentrations using "Base zero" dataset, we investigated whether smaller training datasets could generate efficient ANN models. Hence, two smaller datasets, hereby called "Base A" and "Base B" were constituted for that purpose. "BaseA" dataset comprises

only 32 simulations that test 32 different starting conditions. For each of them, only one input among the sixteen inputs was modified between the two possible values used for "Base zero" (hence a total of 2 × 16=32 starting conditions) while the other inputs were set to the following default initial values: 10 U/mL for enzyme concentration, 1 mM for NADP and 70 mM for both cellobiose and phosphate. "Base B" was built using only 12 different simulation settings. Values for the starting conditions were set as for "Base A" but only E3, E4, E13, $NADP^+$, cellobiose and phosphate initial values were toggled. Hence, "Base B" is a subset of "Base A". E3, E4, E13 were modified to create this "Base B" dataset because they were determined as the enzymes which influence most the output of the system. We also wanted to investigate whether a dataset built by changing only the concentration of these enzymes could be sufficient to identify an adequate model. Henceforth, we would test how efficient is ANN able to learn from few data.

The last dataset, hereafter called "Base V", was generated for the purpose of validating ANN models trained with the small datasets "Base A" and "Base B". This dataset constituted of 625 simulations settings generated after combining different starting concentrations for E3, E4, E13 and NADP. All possible combinations of five initial concentrations for E3, E4, E13 (5, 8.75, 12.5, 16.25, 20 U/mL) and 5 initial concentrations for NADP (0.5, 0.875, 1.25, 1.625, 2 mM) were tested thus accounting a total of 625 starting conditions. All four datasets were obtained through simulations of 8000 min.

The data points for inputs and outputs used in our ANN modeling were all normalized between −1 and 1 with the following equation:

$$X_{norm,i} = \frac{X_i - \frac{X_{max} + X_{min}}{2}}{\frac{X_{max} - X_{min}}{2}}$$

where, X_i is unnormalized data point *i* for an input or an output, X_{min} is the minima among all the data points of an input or output, X_{max} is the maxima among all the data points of an input or output and $X_{norm,i}$ is the normalized value (between −1 and 1) for the data point *i*. Normalization is a recommended pre-treatment of the datasets in neural network modeling. Data are often very heterogeneous in their nature and their range of values. Normalization reduces the dispersion of these values, thus facilitating learning step.

Neural network structure

Artificial neural modeling was performed within the R programming environment using the *caret* and *nnet* packages. The artificial neural network model is a feed-forward neural network with single hidden layer. The works of Cybenko and Funahashi, show that only one hidden layer is sufficient to approximate all non-linear functions [12,13]. Then, the model accuracy depends on the number of the hidden nodes in this layer. The optimal number of hidden neuron was defined automatically by the *caret* package.

Each model had 5 neurons in the hidden layer after the automatic settings. Two distinct artificial neural models were built: one to predict the initial rate of H_2 production and the other for the final concentration of hydrogen. During this fitting step, a 10 fold cross-validation is applied to estimate the performance of the neural network learning phase. The dataset is randomly splitted in 10 parts. Nine parts were used as training dataset and the remaining part was used as the test dataset. An auto-prediction operation is performed too with the identified model for evaluation. During the validation phase, the model is tested on the validation dataset "Base V".

The RMSE (root-mean-square error) and the R^2 (coefficient of determination) were the selected metric in order to evaluate the model performance during the learning phase and the validation:

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{o_i - p_i}{n}}$$

and

$$R2 = \frac{\left(\sum_{i=1}^{n}(o_i - o_m)(p_i - p_m)\right)^2}{\sum_{i=1}^{n}(o_i - o_m)^2 \sum_{i=1}^{n}(p_i - p_m)^2}$$

where,

o_i is the measured activity of the i^{th} data point,

o_m is the mean of the measured activity,

p_i is the predicted activity of the i^{th} data point,

p_m is the mean of the predicted activity,

n is the number of data points,

RMSE was calculated from normalized data between −1, 1.

Results and Discussion

System optimization for final hydrogen concentration

Our knowledge-based model was constructed by following the setup described by Ye et al. and Ardao et al. [5,9]. In these operating conditions, with 70 mM initial cellobiose concentration, the system yielded a final concentration of H_2 of 112 mM after 9000 min of operation time (Figure 2). With this model, several optimization operations was performed in order to search for kinetic parameters that would give the best hydrogen yield after 9000 min with 70 mM of

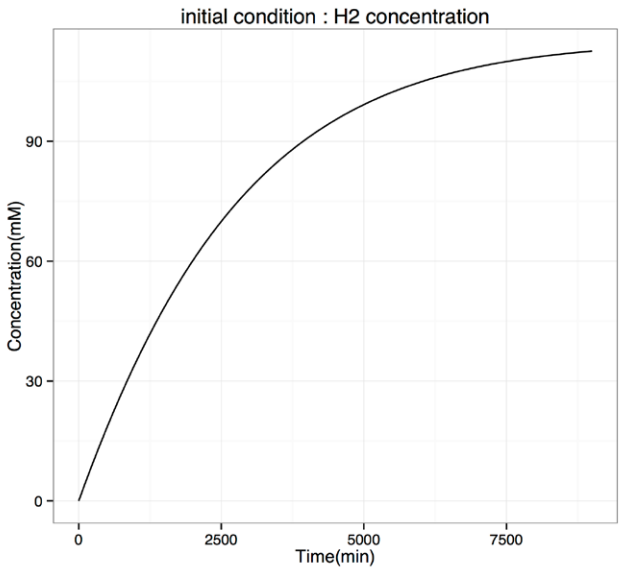

Figure 2: Simulation of hydrogen production in a cell-free system described by Ye et al. [5].

The starting concentrations for cellobiose and inorganic phosphate were 70 mM cellobiose and 1 mM for NADP. The simulated final H2 concentration after 9000 min was 112 mMin agreement to previous reports [5].

cellobiose. Any improvement of final concentration of hydrogen in the same lapse of time would indicate that an improvement of the system productivity is possible if the corresponding parameters could be tweaked. We screened 14 k_{cat}, 19 K_m, 13 enzyme initial concentrations (E_0) and 9 enzyme degradation constants (λ) which had half-life values inferior to 75 h. The advantage of an *in silico* optimization operation is the ease to manipulate all these parameters at the same time and see the repercussion on the simulation. It would require a long and expensive phase of enzyme engineering to do the same operation *in vitro*. It will be easier to validate and reproduce *in vitro* the result from an *in silico* optimization if the number of optimization parameters is low.

First, single stage optimizations were performed as described in the Method section. Four distinct tracts of single stage optimizations were performed, one for each parameter (E_0, k_{cat}, K_m, λ). Hence, multiple single stage optimizations were executed individually for each enzyme to explore the range of values taken by the four investigated parameters after optimization. This range of values is the consequence of the utilization of genetic algorithm which allows finding not the best solution but a set of solutions near the optimal performance of the system. In Table 1 are featured the best final hydrogen concentrations that could be reached after optimizing each of the four parameters individually. These results show that enzyme degradation rate (λ) has no impact on final hydrogen concentration produced, i.e 118 mM, at least within the range of values that was studied. On the other hand, our results show that final H_2 yield improved 2-fold with K_m optimization, 5 fold with E_0 and 7-fold with k_{cat}. As expected, it is the k_{cat}, which is the turnover rate of the enzyme, which had the greater influence on the yield of the system.

Our results show that, with a starting concentration of 70 mM cellobiose, the system can produce up to 840 mM of H_2 if the enzymes are changed for their k_{cat} values. In practice, this can be achieved, likewise for K_m values, either by using better performing homologous enzymes or by protein engineering. The optimization conditions on k_{cat} reached nearly 96% of the theoretical maximum. The optimization with E_0 is second best and H_2 production reached 70% of theoretical maximum, at least within the range of values studied (see Method section). The definite advantage with the optimization of E_0 is that it is easiest to reproduce experimentally. This can be achieved by simply adjusting the starting amount of enzymes to the optimal E_0 values found through optimization. The knowledge of optimal conditions for E_0 is hence a great aid for aiding in implementing such systems.

The knowledge of rate limiting enzymes in the system is key to fully characterize such synthetic systems. Classically, these are established based on their k_{cat} values. How these rate limiting enzymes in combination with the other enzymes influence the productivity of the system is difficult to assess in the wet lab. This assessment proceeds mainly by testing various conditions through a trial and error strategy or more rationally design experiments. Previous published data by Ye et al. [5] and Ardao et al. [9] showed that enzyme E13 is the main limiting factor in the system because it is the enzyme directly implicated in hydrogen synthesis (Supplementary Table 1). To a lesser degree, enzymes E1, E2, E3 and E4 were also declared as limiting enzyme.

We wanted to re-evaluate how these enzymes are key towards productivity of this synthetic system. What simulation and optimization advantageously allow is to test millions of different conditions *in-silico* that is inaccessible experimentally and give access to the range of values for every parameter that allows near-optimal productivity of the system. The analysis of these range of values showed that E3, E4 and E13 enzymes had more impact on the productivity. Indeed, their parameters allowed a near-optimal productivity in a very narrow range of values (Figure 3). Thus, this indicates that, beyond these values, the productivity would be sub-optimal. For the other 10 enzymes, variations in their kinetic properties impacted less on the productivity of the system, for their values could fluctuate with greater amplitude while the system remained close to optimal productivity. To further provide evidence for the importance of these three enzymes, single stage optimization was re-run using only the parameters coming from E3, E4 and E13. The near-optimal H_2 production that could be reached when E_0, K_m and k_{cat} for these three enzymes were optimized is shown in Table 2. They are in very good agreement with those featured in Table 1 thus indicating that these three enzymes are indeed key enzymes for the productivity of the system. Though Ardao

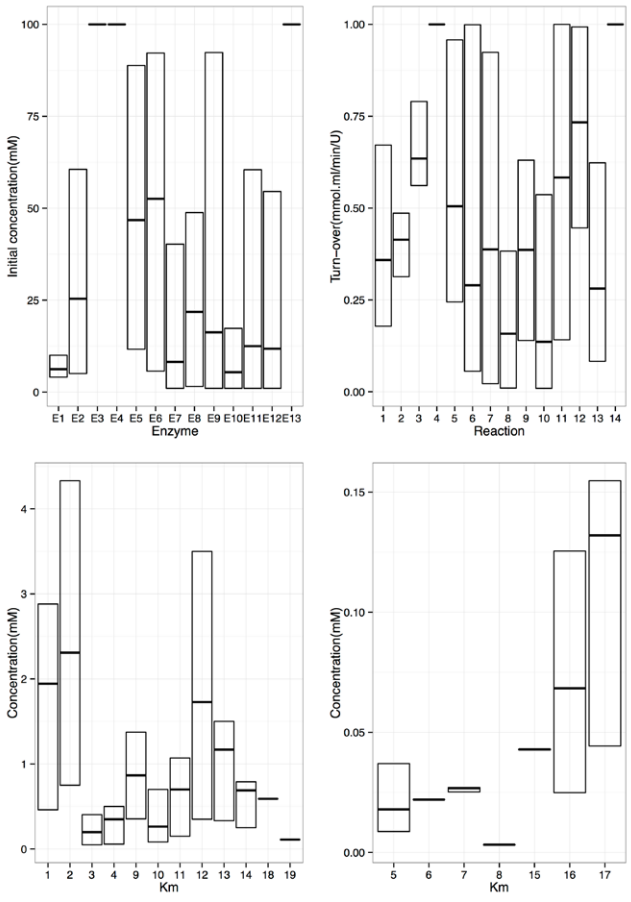

Figure 3: Distribution of optimal values for E_0, kcat and K_m obtained from multiple single stage optimizations.

Under these conditions, near-optimal H2 production was obtained. Horizontal bars indicate average values. Top left shows results after optimization of E0 for all 13 enzymes; cellobiosephosphorylase (E1), phosphoglucomutase (E2),glucose-6-phosphate dehydrogenase (E3), 6-phosphogluconic dehydrogenase (E4), ribose 5-phosphate isomerase (E5),ribulose-5-phosphate 3-epimerase (E6), transketolase (E7), transaldolase (E8), triose-phosphate isomerase (E9), aldolase (E10), fructose-1,6-bisphosphatase (E11), phosphoglucoseIsomerase (E12), $NADP^+$dependent hydrogen dehydrogenase (E13). Top right shows results after optimization of kcat for all 14 reactions (see Method and Supplementary Table 1 for detailed description of the reactions); Bottom are the results after optimization of Km for all 19 substrates (1:$Km_{Pi}$of reaction r1; 2:Km_{Cb} of r1; 2:Km_{Cb} of r2; 3:Km_{G1P} of r2; 4:Km_{G6P} of r2; 5:$Km_{G6P}$of r3; 6:Km_{NADP} of r3; 7:Km_{6PG} of r4; 8:Km_{NADP} of r4; 9:Km_{Ru5P} of r5; 10:Km_{Ru5P} of r6; 11:Km_{DHAP} of r9; 12:Km_{g3P} of r9; 13:Km_{DHAP} of r10; 14:Km_{g3P} of r10; 15:Km_{FdP} of r11; 16:$Km_{F6P}$of r12; 17:Km_{G6P} of r12; 18:Km_{NADP} of r13; 19:Km_{NADPH} of r13).

et al. [9] had shown the importance of these enzymes but only within the framework of E_0 optimization, to our knowledge, this is the first time that this is documented for the Ye et al. [5] synthetic system after optimization of E_0, K_m and k_{cat}. The biology of these enzymes tells us that indeed their activities influences directly hydrogen production: E13 (NADP$^+$ dependent hydrogen dehydrogenase) catalyzes the main reaction in charge of hydrogen production using NADPH, while E3 (glucose-6-phosphate dehydrogenase) and E4 (6-phosphogluconic dehydrogenase) are the dehydrogenases that produce NADPH needed by E13.

We further applied the double stage optimization protocol described in Method section whereby, consecutive to a single stage optimization on a parameter and after retaining its optimal values for all the enzymes, a second optimization step was performed on a second parameter. This should, in principle, improve the final H_2 concentration when compared to a single stage optimization. After a first optimization on E_0, a second optimization was performed on K_m and λ separately, either on all enzymes or only on enzymes E4 and E13. As shown in Table 3, these double optimization drastically augmented the final H_2 concentration compared to single stage optimizations performed on E_0, K_m and λ (Tables 1 and 2). Although the optimization of λ alone was ineffective, the coupling E_0+λ optimizations allowed having almost 95% of the maximum theoretical H_2 production. The E_0+K_m double stage optimization showed that there are sets of values for these two parameters whereby the system can achieved about 98% maximum productivity. The optimal values found for these parameters are basically theoretical values. As discussed above, it is quite easy in practice to modulate E_0 experimentally towards their derived optimal values. However, changing the affinity of an enzyme for its substrate or improving its thermodynamic stability is more complicated and requires better performing homologous or engineered enzymes.

Analysis of kinetics of hydrogen production in optimized systems

The previous optimization procedure was executed to improve the final concentration of H_2 after 9000 min operation. We analyzed the kinetics of H_2 production for each of the optimization protocol that we have investigated, whether it was the single stage optimization or double stage optimization. Indeed, the velocity at which final concentration of hydrogen is reached in the system is an important criteria in terms of productivity. The kinetics of H_2 production for optimal conditions found through the four single stage optimization protocols and through the "E_0+K_m" and "E_0+λ" double stage optimization protocols are given in Figure 4. As expected, the best kinetics was obtained by optimizing k_{cat}. The initial rate calculated from slope at t=0 (Figure 4) is 2,6 mmol.L^{-1}.min^{-1} H_2. Though double stage optimizations on E_0+K_m or E_0+ λ had similar final H_2 yields as single stage optimization on k_{cat}, i.e final concentration of H_2 of about 800 mM (Figure 4, Tables 1 and 3), their kinetics were however different. For short production runs, inferior to 2000 min, optimization by k_{cat} is more advantageous (Figure 4). For longer times of operations, both single stage and double stage optimizations will yield almost equivalent final concentrations of hydrogen.

Our *in silico* optimizations hence allowed us to find the theoretical ideal conditions (i.e values of E_0, K_m, k_{cat} and λ for all enzymes) that optimizes the synthetic system for production of hydrogen. Our results clearly highlight which enzymes must be improved in order to achieve better performance of the system. These are interestingly E3 (glucose-6-phosphate dehydrogenase), E4 (6-phosphogluconic dehydrogenase) and as expected, E13 (NADP$^+$ dependent hydrogen dehydrogenase). Improvement of their performance can be obtained either by finding better performing homologous enzymes from the biodiversity or through enzyme engineering.

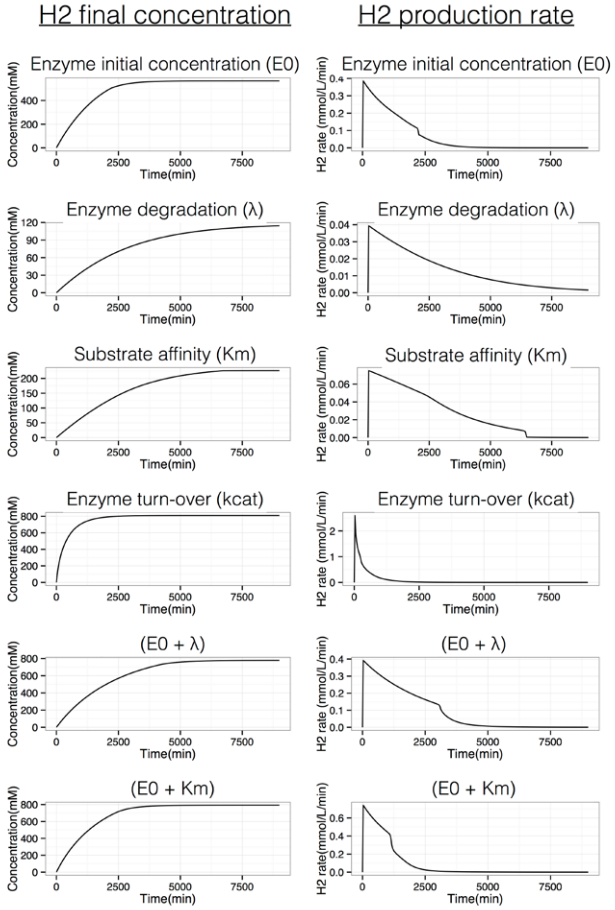

Figure 4: Evolution of hydrogen concentrations (left) and rates of H2 production (right) using optimal values of the parameters that were optimized by single or double stage optimization protocols.

The parameters that were optimized by single stage optimization are enzyme initial concentration (E_0), enzyme degradation parameter (λ), enzyme substrate affinity (K_m) and enzyme turn-over (kcat). The combination of parameters that were optimized by double stage optimization was E_0+λ and E_0+K_m. Simulations time were of 9000 min with 70 mM cellobiose, 70 mM of phosphate and 1mM of NADP as starting substrate concentrations.

An online ODE simulator for the cell-free system

We have developed a web-based simulator that implements the ODE model for the Ye et al. [5] cell-free system. The simulator is freely available at http://www.bo-protscience.fr/h2/ for non-commercial usage. Users can freely tweak with the kinetic parameters and initial E_0 and substrate concentrations. The kinetic laws and the default values for the kinetic parameters are those provided in Supplementary Table 2. Default conditions for the initial concentrations are those indicated in the Method section. The user is provided with a customizable result page: an interactive Google Charts graphic displaying the evolution of the concentration of a substrate, an intermediate metabolite or of the product is displayed. The user is provided with the option to choose which compounds to display. It is also possible to download the whole

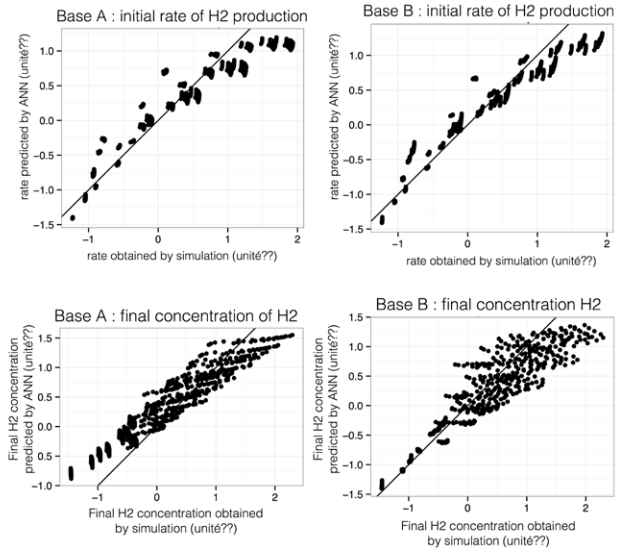

Figure 5: Performance of ANN models to predict the initial rate of production of H2 (top) and final hydrogen concentration (bottom) for 625 data points from the validation dataset.

The models were built on input/output vectors from 32 simulations (BaseA) and 12 simulations (BaseB). Straight line on each graph represents bisector. Values on *x* and *y* axis are normalized values for both rate of H2 production and final H2 concentration.

	Initial rate of H_2 production (kinetics model)		Final H_2 concentration (yield model)	
	RMSE	**R^2**	**RMSE**	**R^2**
Autoprediction	0.006	0.99	0.031	0.996
***k*-fold cross-validation (*k*=10)**	0.006	1	0.043	0.992
Validation	0.202	0.952	0.324	0.905

Table 4: Performance of ANN models built on the "Base zero" dataset. This learning dataset is composed of input and output vectors derived from 65,536 different simulations of the H2 producing system. Shown are the Root Mean Square Errors (RMSE) and the coefficient of determination (R2) of the kinetics and yield models. Results for autoprediction and k-fold cross validation were evaluated on the learning dataset itself. Results for validation were evaluated on the independent "Base V" dataset (see Method for details).

simulation results in the form of a flat text file containing comma separated values of concentration for all metabolites as a function of time. This simulator for example can be useful to test very simply how the improvement of an enzyme would impact the H_2 production.

Using artificial neural networks to model H_2 productivity of the cell-free system, two ANN models were built (see Methods section), one for predicting the rate of H_2 production and another for predicting the final H_2 concentration. Results for these two models built after the "Base zero" dataset which is derived from 65,536 simulations of H_2 production are given in Table 4. These results show that, both ANN models which had a 5-neurons hidden layer were able to predict accurately the kinetics and the yield of the synthetic system. This shows that, provided a database of H_2 production trajectories is available, it is possible to model the output of this H_2 producing synthetic system with the prior knowledge of the starting conditions, i.e the inputs.

In real life situations, it is impossible to generate experimentally such a large database because of cost and time constraints (for 65,536 runs of the system each of 9,000 minutes, 409,600 days would be required and total would cost more than 10 million USD). In fact, we wanted to approach real life situations by building smaller training datasets following experimental design principles and check whether these are sufficient to generate good models using ANN. We hence investigated the ability of ANN to build good models with two much smaller training datasets. The first dataset, "Base A", was built from 32 simulations and the second dataset, "Base B", was derived from 12 simulations. After training with these two datasets, we evaluated how the ANN models accurately predicted the rate of production and the final concentration of H_2. Detailed results given in Table 5 show indeed that, with as few as 12 and 32 inputs/outputs in the training set, the ANN models can remarkably predict with good accuracy both the kinetics and yield of the H_2 producing system with RMSE < 0.36 and R^2>0.86. Figure 5 shows the corresponding scatter plots for the "Base V" validation dataset which comprises 625 data points. The data were all normalized between 1, −1 prior to the learning phase (see Method). Interestingly, the models predicted values superior to 1 and inferior to −1. The validation dataset contained data points that were indeed beyond these boundaries but the models were less accurate in these areas. Overall, rate of H_2 production was predicted with lower RMSE than for final H_2 concentration (Table 5) and as illustrated by the lesser dispersal of the data points away from the bisector line (Figure 5). Interestingly, "Base A" and "Base B" generated ANN models with very similar RMSE and R^2 values on the validation set and with overall performance close to the more accurate models built with the much larger "Base zero" dataset. This indicates that ANN is quite robust to the size of learning datasets provided that the learning datasets are appropriately designed. It also indicates that restraining the ANN inputs to only the 3 rate limiting enzymes and to the starting substrates (as for "Base B") is indeed an interesting design strategy. Also, it can be concluded that small learning datasets which are easier to construct experimentally can provide very satisfactory models for predicting kinetics and yield of such a synthetic system.

Conclusion

Modeling is relevant to understand, manipulate, and evaluate a biochemical system. From the works of Ye et al. and Ardao et al. [5,9], we have implemented a knowledge-based model under the SBML format. System metabolites and the interaction between them during reactions were numerically encoded in ODE which allowed us to perform *in silico* optimizations and to provide answers as for how to improve of hydrogen production. We stated different system modifications that would yield higher hydrogen productivity. Among these, the manipulation of the initial concentration of enzyme (E_0) is the easiest to reproduce experimentally. The manipulation on the k_{cat} is best to achieve optimal performance of the system. Optimization on K_m and enzyme degradation constants can be looked for but they had less impact on the system.

We studied double stage optimization combining E_0 with K_m or λ. The advantage of these double stages optimizations is that they can be easily applied experimentally. As it is difficult to find or engineer enzymes with improved k_{cat}, this double stage optimization involving E_0 in first stage and K_m or λ optimization in second stage is a very good alternative.

In the same line, we evaluated the most pertinent set of enzymes for optimization so as to reduce the combinatorial complexity of the experimental design. Three enzymes, E3 (glucose-6-phosphate dehydrogenase), E4 (6-phosphogluconic dehydrogenase) and E13

	Initial rate of H_2 production (kinetics model)				Final H_2 concentration (yield model)			
	RMSE		R^2		RMSE		R^2	
	Base A	Base B	Base A	Base B	Base A	Base B	Base A	Base B
Autoprediction	0.003	0.001	1	1	0.15	0.001	0.8	1
***k*-fold cross-validation (*k*=10)**	0.086	0.124	0.791	1	0.237	0.135	0.661	1
Validation	0.308	0.265	0.882	0.91	0.358	0.353	0.92	0.86

Table 5: Performance of ANN models built on the "Base A" and "Base B" datasets. These were composed of input and output vectors derived from respectively 32 and 12 different simulations of the H2 producing system. Shown are the Root Mean Square Errors (RMSE) and the coefficient of determination (R2) of the kinetics and yield models. Results for autoprediction and *k*-fold cross validation were evaluated on the learning datasets themselves. Results for validation were evaluated on the independent "Base V" dataset (see Method for details).

($NADP^+$ dependent hydrogen dehydrogenase) that were limiting for H_2 production, were identified through our simulations. E3 and E4 are indeed producing NADPH that is required by E13 which is directly involved in H_2 production. Restraining the optimization to these three enzymes showed to yield near-optimal H_2 production.

All these results provide new insights into the properties of the cell-free system of Ye et al. [5]. Though, this study is based on simulations. The effectiveness of the method needs further validation using real experimental data. Other cell-free systems for hydrogen production were recently designed [14,15]. It could be interesting to likewise apply on these systems the same modeling and optimization approaches we used here. But ODE models require the knowledge of kinetic parameters and laws for these new enzymatic systems. These information are not always fully available and may remain unknown for long. We explored a modeling approach by ANN to have a global overview of the performance of a synthetic system through the sole knowledge of few initial enzyme and substrate concentrations and the corresponding outputs. This approach requires hence only empirical data. We used the cell-free system of Ye et al. [1] as object of modeling and our knowledge-based ODE model as generator of empirical data. We showed that ANN model with practical dataset sizes allowed accurate prediction of the kinetics and yields of the system. The knowledge of the limiting enzymes was interesting since it allows simplification of the experimental design for constituting the learning database. Our ANN model works better to predict the initial rate of H_2 production than the final concentration of hydrogen. Further works are needed to elaborate a way to resolve this case.

ANN approach hence gives very basic information about the kinetics of the system but it can still be useful meanwhile more accurate models like ODE models are made available. Further, since the inputs of ANN are the initial concentrations of the starting substrates and enzyme, it is also conceivable to build an optimization protocol to search for ANN-derived optimal starting conditions. This study is solely based on simulations. The effectiveness of the methods need further validation using real experimental data. In particular, it is at stake to demonstrate experimentally how higher yields could be achieved by substituting key enzymes identified in this study by better performing versions.

Acknowledgement

NF is supported by a PhD grant from Conseil Régional de La Réunion and European Union (Fonds Social Européen) under file number 20112717-tiers 157173. BGP and FC have been supported by Conseil Régionale de La Réunion. BO acknowledges support from Université de Nantes. The authors thank Peaccel SAS for providing infrastructural and technical support for hosting the bioH2 simulator.

References

1. Eroglu E, Melis A (2011) Photobiological hydrogen production: Recent advances and state of the art. Bioresour Technol 102: 8403-8413.
2. Oh YK, Raj SM, Jung GY, Park S (2011) Current status of the metabolic engineering of microorganisms for biohydrogen production. Bioresour Technol 102: 8357-8367.
3. Hallenbeck PC, Abo-Hashesh M, Ghosh D (2012) Strategies for improving biological hydrogen production. Bioresour Technol 110: 1-9.
4. Zhang YH (2010) Production of biocommodities and bioelectricity by cell-free synthetic enzymatic pathway biotransformations: challenges and opportunities. Biotechnol Bioeng 105: 663-677.
5. Ye X, Wang Y, Hopkins RC, Adams MW, Evans BR, et al. (2009) Spontaneous high-yield production of hydrogen from cellulosic materials and water catalyzed by enzyme cocktails. ChemSus Chem 2: 149-152.
6. Woodward J, Orr M, Cordray K, Greenbaum E (2000) Enzymatic production of biohydrogen. Nature 405: 1014-1015.
7. Zhang YH, Evans BR, Mielenz JR, Hopkins RC, Adams MW (2007) High-yield hydrogen production from starch and water by a synthetic enzymatic pathway. PLoS One 2: e456.
8. Zhang YHP (2009) A sweet out-of-the-box solution to the hydrogen economy: is the sugar-powered car science fiction? Energy Environ Sci 2: 272-282.
9. Ardao I, Zeng AP (2013) In silico evaluation of a complex multi-enzymatic system using one-pot and modular approaches: Application to the high-yield production of hydrogen from a synthetic metabolic pathway. Chem Eng Sci 87: 183-193.
10. Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, et al. (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 19: 524-531.
11. Matsumoto M, Nishimura T (1998) Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. ACM Trans Model Comput Simul 8: 3-30.
12. Cybenko G (1989) Approximation by superpositions of a sigmoidal function. Mathematics of control, signals and systems 2: 303-314.
13. Funahashi K (1989) On the approximate realization of continuous mappings by neural networks. Neural Networks 2: 183-192.
14. Martín del Campo JS, Rollin J, Myung S, Chun Y, Chandrayan S, et al. (2013) High-yield production of dihydrogen from xylose by using a synthetic enzyme cascade in a cell-free system. Angew Chem Int Ed Engl 52: 4587-4590.
15. Myung S, Rollin J, You C, Sun F, Chandrayan S, et al. (2014) In vitro metabolic engineering of hydrogen production at theoretical yield from sucrose. Metab Eng 24: 70-77.