ARTICLE

BIOTECHNOLOGY and BIOENGINEERING

Coping With Complexity: Machine Learning Optimization of Cell-Free Protein Synthesis

Filippo Caschera,[1,2] Mark A. Bedau,[1,3] Andrew Buchanan,[1] James Cawse,[1,4] Davide de Lucrezia,[5,6] Gianluca Gazzola,[1,7] Martin M. Hanczyc,[1,2] Norman H. Packard[1,5,8]

[1]ProtoLife, Inc., 57 Post St #513, San Francisco, California 94104;
e-mail: n@protolife.com; martin@ifk.sdu.dk
[2]Department of Physics and Chemistry, University of Southern Denmark, Odense, Denmark
[3]Reed College, Portland, Oregon
[4]Cawse and Effect, Pittsfield, Massachusetts
[5]European Center for Living Technology, Venice, Italy
[6]Explora S.r.l., Roma, Italy
[7]Rutgers Center for Operations Research, Rutgers University, New Brunswick, New Jersey
[8]Santa Fe Institute, Santa Fe, New Mexico

Received 2 November 2010; revision received 29 March 2011; accepted 4 April 2011
Published online in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/bit.23178

ABSTRACT: Biological systems contain complex metabolic pathways with many nonlinearities and synergies that make them difficult to predict from first principles. Protein synthesis is a canonical example of such a pathway. Here we show how cell-free protein synthesis may be improved through a series of iterated high-throughput experiments guided by a machine-learning algorithm implementing a form of evolutionary design of experiments (Evo-DoE). The algorithm predicts fruitful experiments from statistical models of the previous experimental results, combined with stochastic exploration of the experimental space. The desired experimental response, or evolutionary fitness, was defined as the yield of the target product, and new experimental conditions were discovered to have ~350% greater yield than the standard. An analysis of the best experimental conditions discovered indicates that there are two distinct classes of kinetics, thus showing how our evolutionary design of experiments is capable of significant innovation, as well as gradual improvement.

Biotechnol. Bioeng. 2011;xxx: xxx–xxx.

KEYWORDS: design of experiments; protein expression; complexity; optimization

Introduction

The pioneering work of Nirenberg and Matthaei (1961) demonstrated that cell-free protein synthesis is possible, that is, synthesis of proteins without the entire machinery of a living cell. Since then, considerable progress has been made in refining cell-free systems (Katzen et al., 2005; Zubay, 1973), partly toward a better understanding (Underwood et al., 2005), and partly toward increasing productivity. The use of in vitro translation systems can have advantages over in vivo gene expression when the over-expressed product is toxic to the host cell, when the product is insoluble or forms inclusion bodies, or when the protein undergoes rapid proteolytic degradation by intracellular proteases. In vitro synthesis of proteins in cell-free extracts is an important tool for molecular biologists and has a variety of applications, including the rapid identification of gene products (e.g., proteomics; Goshima et al., 2008; Sitaraman and Chatterjee, 2009; Woodrow et al., 2006), mutational studies (Chambers et al., 1993; Iffland et al., 2005), protein folding studies (Jiang et al., 2008; Liguori et al., 2008), incorporation of modified or unnatural amino acids for functional studies (Cornish et al., 1994; Klarmann et al., 2007), and investigation of protein–protein (Abe et al., 2007; He and Wang, 2007; Oyama et al., 2006) and antibody epitope mapping (Delbecq et al., 2006; Osada et al., 2009). Cell-free systems for in vitro gene expression and protein synthesis have been described for different prokaryotic and eukaryotic systems (Anderson et al., 1983; Pelham and Jackson, 1976). The most frequently used cell-free translation systems consist of extracts from rabbit reticulocytes, wheat germ,

Contributions: FC, MAB, JNC, GG, MMH, and NHP designed research; FC, AB, GG, and MMH performed research; DDL and FC contributed new reagents or analytic tools; FC, GG, and MMH analyzed data; FC, MAB, DDL, JNC, GG, MMH, and NHP wrote the paper.
Correspondence to: M.M. Hanczyc and N.H. Packard

and *E. coli*. Despite recent advancement, in vitro protein synthesis is severely impaired by low yield and reproducibility; for example, standard in vitro systems produce picomole (or nanogram) amounts of protein per 50 μL reaction. This yield is sufficient for some types of analyses, such as polyacrylamide gel separation, Western blotting, immunoprecipitation and, depending on the protein of interest, enzymatic or biological activity assays. However, the industry standard to produce large-scale amounts of protein relies on large reactors containing bacteria that overexpress the protein of interest. Recently, a number of methods have been developed to increase overall yield of in vitro protein expression to a preparative scale and to improve protocol reproducibility (Forstner et al., 2007; Kim et al., 1996; Ozawa et al., 2004; Spirin et al., 1998; Wang et al., 2008).

Although in vitro protein synthesis systems are remarkably simpler than whole-cell expression systems, they rely on a complex network of interacting factors (i.e., 70S or 80S ribosomes, tRNAs, aminoacyl-tRNA synthetases, initiation, elongation and termination factors, amino acids, energy sources, energy regenerating systems, and other co-factors) that define a high-dimension compositional space. Sequential adjustment of single factors cannot be effective to optimize the synthesis process (e.g., total yield) since the presence of strongly nonlinear interactions among factors require simultaneous adjustment of parameters. Strong, nonlinear interactions between factors complicate our understanding of the system, and therefore prevent any optimization using models derived from first principles.

Here we demonstrate that an intelligent evolutionary search procedure, Evo-DoE (Caschera et al., 2010; Theis et al., 2006), using a combination of statistical modeling and intelligent stochastic exploration, can make significant improvements in target protein yield. After eight iterations of sparse but intelligent sampling of the high dimensional experimental space, we see ~350% greater yield over the standard. We also observe that our best experiments show two distinct classes of kinetics that underlie the improved response of the in vitro protein synthesis system.

Materials and Methods

Reagents

Water DNase RNase free, hydrochloric acid 37%, HEPES buffer, potassium hydroxide, all the 20 standard canonical amino acids in powder, phosphoenolpyruvate (PEP), glucose, β-nicotinamide adenine dinucleotide (NAD), and magnesium chloride were purchased from Sigma-Aldrich (Milan, Italy). The stock solutions of amino acids, PEP, glucose, β-nicotinamide adenine dinucleotide, and magnesium chloride were stored in Eppendorf tubes 1.5 mL at −20°C after preparation and thawed before performing each experiment. The enhanced green fluorescent protein (eGFP) used to calibrate the experiment using a standard curve was purchased from Biosensis (Thebarton, South Australia). The lyophilized DNA library was synthesized by Explora S.r.l. (Rome, Italy) and stored at −20°C after rehydration with water. The Expressway™ Maxi Cell-Free *E. coli* Expression System was purchased from Invitrogen, Milan. The robot workstation used Rainin tips GPS-L250 SPACE SAVER 960PZ purchased from Elettrofor S.a.s. (Rovigo, Italy). The reactions for protein synthesis were performed in 384-well plate black purchased from Sigma-Aldrich.

Solutions

A 50 mM HEPES solution was prepared at pH 7.5 with a solution 5 M potassium hydroxide. PEP, glucose, nicotinamide adenine dinucleotide, magnesium chloride, and glucose powders were dissolved in HEPES 50 mM pH 7.5 at the desired concentration. Separate stock solutions at 3 M were made for each amino acid. The amino acids mixture minus methionine, needed for the synthesis reaction, was prepared by mixing together 17 μL from each 3 M stock and then diluted with 415 μL of 5 M KOH, achieving a clear solution with final amino acid concentration of 69 mM each. The solution was diluted with HEPES 50 mM, pH 12, KOH at the concentrations required in the experiment. Methionine was dissolved in HEPES 50 mM, pH 12, KOH at the desired concentrations. The cellular extract was thawed and diluted 1:1 with water.

High-Throughput Protocol

The high-throughput experiment was performed with a robotic workstation for liquid handling, Xiril 75-1-2 (Hombrechtikon, Switzerland). The hardware layout used was assembled specifically for the experimental protocol used. The Eppendorf tubes 1.5 mL that contain the stock solutions were set in 32 position racks for 1.5 or 2 mL microfuge tube with lids. The racks were purchased from Xiril. Two premixes with different concentrations of the ingredients PEP, glucose, magnesium chloride, and nicotinamide adenine dinucleotide (NAD) were made as follows. From concentrate solutions prepared in 1.5 mL Eppendorf tubes, PEP (0.8, 0.6, and 0.3 M), glucose (1.6, 0.8, and 0.4 M), magnesium chloride (0.8, 0.6, and 0.3 M), and NAD (0.03, 0.02, and 0.009 M) were dispensed 20 μL each into the target positions in a 384-well plate according to a well-map producing the premixes at the desired concentrations. After preparation of the premix (40 μL per well), the plate is removed and stored at 4°C. Twenty-four microliters of reaction buffer, 24 μL of diluted *E. coli* cellular extract, 5 μL of amino acid mix, 5 μL of methionine, 5 μL of plasmid DNA and 1.2 μL of T7 polymerase solution were dispensed according to the well-map. The distribution of all solutions was handled by the robotic workstation except for the T7 polymerase. All the sources of stock solutions were

contained in 1.5 mL Eppendorf tubes set in 32 position racks for 1.5 mL tubes.

The fluorescence intensity of the reaction mixtures was measured ($t = 0$ min) and the reactions were then incubated in a thermoshaker at 37°C for 30 min. A second fluorescence reading was taken ($t = 30$ min) and the well-plate was placed again in the robotic work station. Thirty microliters feed buffer, 5 μL amino acid mix, 5 μL methionine solution, 9 μL from PEP–glucose premixes and 9 μL from magnesium chloride–NAD premixes were added and the well-plate was then returned to the thermoshaker at 37°C to continue the reaction. Additional time points were taken by fluorescent measurement at 90 and 120 min. After this the well-plate was put again in the robotic workstation. Nine microliters PEP–glucose premixes and 9 μL magnesium–NAD premixes were added to the reaction mixtures according to the well-map. eGFP synthesis by fluorescence was then measured at $t = 180$, 240, 300, and 360 min. To avoid undesired noise in protein synthesis efficiency due to multiple cycles of freezing and thawing of reagents in each experiment (generation), the solutions of *E. coli* cellular extract, reaction buffer and feed buffer stored at −80°C were thawed only once and then immediately used in the experiment. The same lot of cellular extract was used for all the experiments.

Modeling Protocol

In choosing random samples throughout all generations, the probability distribution over the experimental space was shaped to make choices near experiments already chosen less probable. Specifically, the probability of an experiment to be sampled was proportional to the Euclidean distance between such experiment and the closest already sampled one. The probability distribution was recomputed after sampling every experiment, and a previously sampled experiment could not be resampled. Predictions of experiments were obtained from a neural network model (learned with back-propagation using *nnet* in the R language after standardizing all inputs and normalizing the output to the [0,1] interval) with 16 inputs and one output (each DNA type was regarded as a separate dimension, taking on either zero or one, with the constraint that only one of these dimensions could be non-zero, yielding a 16-dimensional input space). Each neural network was constructed with particular metaparameter values (weight decay and number of hidden nodes). The model's metaparameters were selected using a bagging process (Breiman, 1996), repeating the model learning on 40 different data sets, each being a different random sample of 90% of the observed experiments, and 10 times on each data set. Each configuration of metaparameters was then assigned a quality measure, calculated as the median correlation between the remaining 10% observations and the corresponding predictions over all the repeats. Predicted experiments for generations two through four used a predicted-fitness-proportional sampling criterion, which consisted in predicting the fitness of all unobserved experiments (from the total of 1,572,864 experiments in the space) and then sampling them with a probability proportional to their predicted fitness; generation five was created with the same criterion applied after raising the predicted fitnesses to the 2nd power in order to bias the choice toward higher fitness; generation 6 was created with the same criterion as generation 5, but raising the predicted fitnesses to the 4th power; generation seven was created applying the predicted (raw) fitness-proportional criterion, used for generations two to four, but only to the predicted top 1/1,000 experiments; the last generation was created with the same criterion as generation seven, but applied to the predicted top 1/10,000 experiments.

Results and Discussion

We begin with an established cell-free protein synthesis system and aim to increase the system's yield through a process of optimization. The cell-free system we use as a standard is available as a kit from Invitrogen, based on developments reported in the literature (Kim et al., 1996). It uses a cell extract from *E. coli* (Zubay, 1973), which contains all the cellular components needed for transcription and translation. In addition, it is necessary to feed the molecular machineries involved in the transcription and translation processes with a continuous flow of fresh ATP (Calhoun and Swartz, 2007; Kim and Swartz, 1999). This, in turn, requires the use of several enzymes to produce a cell-free metabolism with all the relevant pathways derived from prokaryotic cells, with the addition of an ATP regenerating system (Calhoun and Swartz, 2007; Kim and Swartz, 1999, 2001; Kim et al., 2006; Lesley et al., 1991; Studier et al., 1990). The target protein is eGFP, and the yield is measured by a fluorometric assay. This assay measures the presence of only fully functional (folded) protein, so that the process of refolding and protein maturation is part of the optimization. The observed fluorometric signal defines a response surface over a space of experimental variables that are varied during the optimization process. The experimental response function, or fitness function f, was defined as $f = \max_t(F^t - F^t_b)$, where F^t is the fluorescence of the experiment at time t, and F^t_b is the fluorescence of a blank well at the same time.

In our iterated high-throughput experiments, each generation is a set of experiments that takes place in a 384-well plate. In the in vitro protein expression system, there are many parameters that could be varied independently or simultaneously in an attempt to improve protein yield. Also, there is no definitive list of the chemical components present in the *E. coli* cell extract. This implies a design of experiments problem with an extremely high-dimensional experimental space and noise due to unspecified components from the extract. A drastic reduction of the experimental space is obtained by fixing the relative ratios of all the amino acids, except for methionine. Also, we diluted one of the more expensive components of the system, the *E. coli* extract, on the order of 1:1. This was done (1) to create a suboptimal condition that our system could

then optimize (2) to enable more experiments using less reagent, and (3) to help minimize the cost of each experiment.

Table I lists the components of our system that we optimized, as well as the experimental space. The system is divided into four modules, each added at different times. Each module is made of two or more components, and its composition may be varied during the optimization process. The protocol starts with module one, at time $t = 0$. At $t = 30$ min, module two is added. Then modules three and four are added according to two variable timing parameters, either at $t = 30$ min or at $t = 120$ min. All values specify concentration (mM), except *DNA*, which varies between six different constructs as described below, and the final two parameters, which are the times that module 3 and module 4 were added. Bold values define the standard. The final column indicates the short name used for each experimental coordinate. A particular experiment is obtained by specifying values for each of the variable components, that is, by specifying a vector of values (DNA, AA_0, AA_{30}, M_0, M_{30}, PEP, Mg, G, NAD, T_1, T_2). Thus the experimental space may be considered to have 11 dimensions, and the number of possible experiments in the experimental space, after it has been discretized as shown in Table I, is simply the number of possibilities for each variable component multiplied together, or $6 \times 4^8 \times 2^2 = 1{,}572{,}864$. We simultaneously vary three types of parameters: (i) concentrations of particular ingredients, with values shown in mM units, (ii) DNA construct, and (iii) timing parameters that determine when modules three and four are added.

The sequence of the DNA construct greatly influences the functionality of the cell-free system, for instance small terminators sequences for the T7 polymerase are shown to effect the multigene expression in vitro and in vivo (Liping et al., 2009). Cell-free protein expression is a multi-step process involving mRNA synthesis and subsequent translation of the gene of interest (GOI) encoded in a suitable DNA template that also contains the proper context of transcription and translation regulatory elements, including an RNA polymerase promoter, the Shine-Dalgarno ribosome binding site (RBS), an ATG initiation codon, the GOI, a stop codon (TAA), and finally a transcription terminator, as illustrated in Figure 1. Although these regulatory elements and their reciprocal locations are well-characterized, spacer regions (L_1, L_2, and L_3 in number of nucleotides) between regulatory elements may affect transcription and translation yield by influencing enzyme accessibility and processivity during mRNA synthesis and protein expression. The lengths of these three spacer regions are the parameters varied to obtain the six variations of DNA, as described in Table II. All spacers, except for spacer L1 in sequence E, were designed through incremental addition around a common core sequence (Table II) to minimize the influence of nucleotide composition and order on experimental outcomes. Conversely spacer L1 in sequence E was designed to base-pair with the upstream RBS. Other variables related to the DNA and the transcription/translation process, such as mRNA secondary structure and content of the coding region, codon usage, and 5′ and 3′-UTR (untranslated region), may also deeply affect the overall process by influencing messenger stability and accessibility to the translation machinery. These variables are mainly gene-specific and must be optimized for each GOI independently. In contrast, we focus here on the optimization of general regulatory elements common to any DNA template, so that results may be generally applicable to a broad range of target proteins.

Six different types of DNA used were obtained by varying the length and nature of the spacers between the ribosome binding site, the ATG initiation codon, the coding region,

Table I. Description of fixed elements of the protocol, together with those that are varied, defining the experimental space.

	Component	Levels	Name
Experimental components and definition of the experimental space			
Module 1	*E. coli* extract optimized for increased stability of DNA constructs during transcription and translation and increased production of soluble protein	Not varied	
Module 1	Reaction buffer, composed of an ATP regenerating system to provide an energy source for protein synthesis	Not varied	
Module 1	A T7 Enzyme Mix containing T7 RNA polymerase and other components optimized for T7-based expression from DNA templates	Not varied	
Module 2	An optimized feed buffer containing salts and other substrates to replenish components depleted or degraded during protein synthesis	Not varied	
Module 1	DNA sequence variations	**A**, B, C, D, E, F (types)	*DNA*
Module 1	Amino acids (except methionine)	**1.25**, 2.5, 3, 4 mM	AA_0
Module 2		**1.25**, 2.5, 3, 4 mM	AA_{30}
Module 1	Methionine	**1.5**, 3, 3.6, 4.8 mM	M_0
Module 2		**1.5**, 3, 3.6, 4.8 mM	M_{30}
Module 3	Phosphoenolpyruvate (PEP)	**0**, 12, 23, 30 mM	*PEP*
Module 3	$MgCl_2$	**0**,12, 23, 30 mM	*Mg*
Module 4	Glucose	**0**,15, 30, 60 mM	*G*
Module 4	NAD	**0**, 0.32, 0.65, 1 mM	*NAD*
	Time for application of module 3	**30**, 120 min	T_1
	Time for application of module 4	**30**, 120 min	T_2

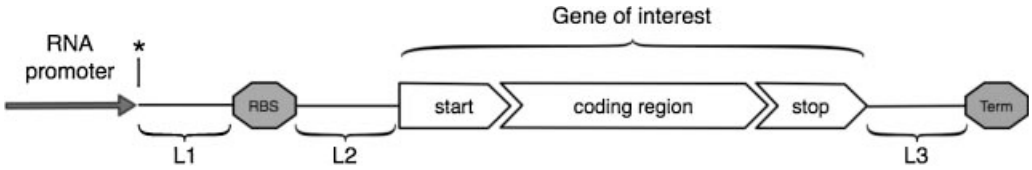

Figure 1. General structure of a DNA template for cell-free protein expression, with the universal regulatory elements for the transcription/translation process: after an RNA polymerase promoter (gray arrow), a ribosome binding site (gray octagon), followed by the gene of interest, and finally the RNA transcription terminator region (gray octagon). L_1, L_2, and L_3 are spacer regions between regulatory elements, and their lengths are varied within the experiment, as described in Table II. The first transcribed nucleotide is marked with a *, just downstream from the RNA polymerase promoter.

and the stop codon. Sequence A was designed according to the guidelines first published by Studier et al. (1990), and represents the standard for the optimization process. We designed sequences from B to D with different spacer region lengths in order to investigate regulatory elements' reciprocal positioning and their influence on overall yield. We designed sequence E with an L_1 spacer complementary to the ribosomal binding site in order to form a stable hairpin structure with RBS. It is expected that this mRNA secondary structure element will influence translation initiation. Finally, we designed sequence F with spacers of suboptimal length, so that the RNA polymerase docking on DNA template and the ribosome minor subunit docking on correspondent mRNA cannot occur simultaneously, thus decoupling the transcription/translation process.

Since the experimental space contains more parameters than those found in the standard, some of the concentration parameters for the standard are zero (in column 3 of Table I), indicating that the relevant component is omitted from the standard. Several results in the literature motivated our use of additional ingredients not found in the standard. For example, there is a potential problem regarding the inorganic phosphate produced during the translation event. If inorganic phosphate is not recycled, the reaction becomes inhibited, because phosphate sequesters the dissolved magnesium ions that are needed for the functionality of the enzymes that catalyze energy production and recycling of waste products (Calhoun and Swartz, 2005). This motivates the addition of magnesium chloride in module three, with variable concentration and time of addition, as indicated in Table I. Another aspect to be considered in designing metabolic reactions for protein production in vitro is the pragmatic value of using an energy source other than ATP. Alternatives that have been investigated include glucose-6-phosphate (Kim et al., 2007a), fructose-1,6-bisphosphate or creatine phosphate (Kim and Swartz, 2000; Kim et al., 2007b), support the addition of PEP in module three. Besides the direct addition of PEP, we allowed the addition of glucose in module four to enable the use of glycolytic reaction pathways to produce ATP, in its interaction with other substances in the undefined composition of the cell extract. Studies focused on understanding the chemical variables that most influence protein yield have reported an important

Table II. Definition of the DNA variants used.

DNA construct	L_1 (bp)	L_2	L_3	Comments
DNA variations				
A	15	7	50	Standard
B	15	10	50	
C	20	7	50	
D	20	10	50	
E	15	7	50	Stable hairpin secondary structure masking RBS
F	5	5	4	Negative standard, spacers of suboptimal length

Space class	Length (bp)	Sequence	Comments
Spacer sequences			
L1	5	GTTTA	
	15	ATTGTTTAACTCTA	
	20	ATCATATTGTTTAACTCTA	
	15	TATCTCCTTCTTTAA	Forms stable hairpin with RBS
L2	5	TACAT	
	7	TATACAT	
	10	TATTATACAT	

See Figure 1 for location of the spacers, L_1, L_2, and L_3. Common core sequences are underlined.

dependence of the reaction productivity on the concentrations of pyruvate, amino acids, and co-enzymes. Significant responses were also observed with the degradation of amino acids and the hydrolysis of the ATP molecules in the absence of protein synthesis, as well the recovered activity when the reaction mixture was supplied with the addition of those components. In addition, the studies showed a positive effect on protein production depending on the concentration of co-factors such as NAD, CoA, and creatine phosphate. These studies are the primary motivation for adding the NAD in module four.

After establishing the space of possible experiments, as described in Table I, we began the optimization process. The first generation consisted of 49 randomly chosen experiments, that is, 49 random choices for a vector of values (DNA, AA_0, AA_{30}, M_0, M_{30}, PEP, Mg, G, NAD, T_1, T_2), along with the standard. In all subsequent generations, random selection of experiments was complemented by selection according to the predictions of the model of fitness across the entire experimental space; this model is built on the data collected in all preceding experiments. The experimental space is so large that the optimization procedure must make a compromise between exploration via random choice and exploitation of structure in the data gathered in previous experiments. Consequently, from the first generation to the final generation, an annealing took place, ranging from pure random choice in the first generation to a choice based exclusively on the predictive model in the last generation. Intermediate generations had a mixture: 5 random and 19 predicted in generations 2–6; 2 random and 19 predicted in generation 7.

Models are commonly used in design of experiments (Caschera et al., 2010), but their use here to provide virtual experimental results and adaptively establish the trade-off between exploration and exploitation is novel. Our use of Evo-DoE is an enhanced version of the techniques used recently in a different optimization experiment (Forlin et al., 2008), and goes beyond the use of standard (Corma et al., 2005) or model-assisted (Stein, 1999) genetic algorithms for evolutionary design of experiments. Our approach is similar to a family of iterative nonlinear optimization techniques that employ kriging models (Jones et al., 1998) to interpolate experimental observations (Cawse et al., 2010). In these models, the response is assumed to be the sum of a polynomial function and a term that represents a systematic deviation from the polynomial. At every iteration, kriging-based optimization techniques extract from the former component of the model a prediction of the global fitness landscape's topology and from the latter component a measure of the prediction's local uncertainty. This information is then used to build up a utility function on the experimental space, whose optima define the sampling criterion for the experiments of the following iteration (Sasena et al., 2005). Evo-DoE differs by employing a combination of two separate stochastic sampling criteria, the one based on the predictions of a more sophisticated class of interpolating model and the other one on a distance matrix built on all previously sampled experiments (as described below).

Each model has inputs that correspond to each of the dimensions of the experimental space, and an output that predicts experimental fitness. The models used were neural networks, with meta-parameters adjusted each generation using a bootstrapping technique described below in the Materials and Methods Section. Once the model for a given generation was built, it was used to predict the fitness of every untried experiment in the space ($\sim 1.5 \times 10^6$ experiments), a version of virtual experimentation that is vastly less time-consuming than conducting real experiments. The next generation of real experiments was determined by sampling this distribution of predicted experimental results, with a bias toward high fitness experiments, as described in the Materials and Methods Section.

The evolutionary progress of the experimental population's fitness is shown in Figure 2. The predictive models became increasingly accurate, starting with a linear correlation of ~20% (between out-of-sample predicted fitness and actual fitness), and ending with a linear correlation of ~90%. Table III shows the explicit specification of the top 12 experiments after the last generation of the evolutionary learning run, as well as the standard (S).

Figure 3 shows evolutionary learning as represented by the level of occurrence of particular values for experimental coordinates within the population of experiments conducted each generation. Significant learning begins around generation 5, even though a significant fitness increase may not be observed until following generations, as illustrated in Figure 2. Changes in the percentage representation are concrete evidence that the predictive evolutionary algorithm is succeeding in its learning task; if no learning were taking

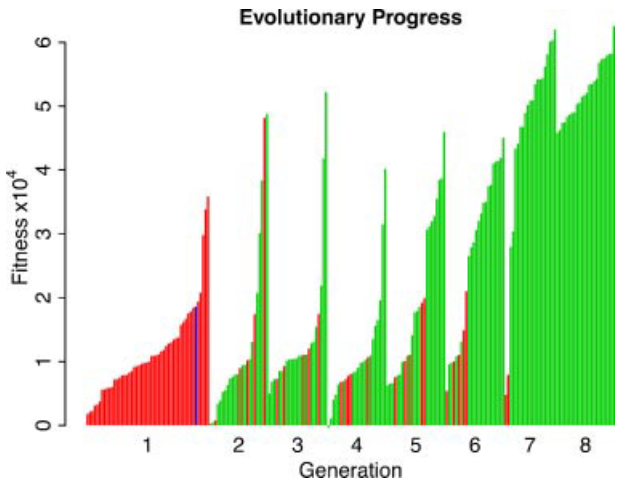

Figure 2. Progress of the predictive evolutionary algorithm over eight generations: experimentally measured fitness as a function of evolutionary time. The standard (averaged over all repeats in each generation) is shown in black. Randomly chosen experiments are shown in gray, experiments from the predictive model in light gray. Fitness is averaged over repeats, when performed.

Table III. Top 12 experiments discovered during the optimization procedure, ordered by fitness value.

Rank	*DNA*	AA_0	AA_{30}	M_0	M_{30}	*PEP*	*Mg*	*G*	*NADH*	T_1	T_2
Top 12 discovered experiments											
1	C	1.25	4	3	3	12	0	0	0	120	—
2	C	1.25	3	3.6	1.5	12	23	15	0	120	120
3	C	1.25	3	3.6	3.6	12	23	0	0.32	30	120
4	C	1.25	4	3	1.5	23	0	15	0.32	120	120
5	C	1.25	4	3.6	4.8	30	23	15	0.32	120	120
S	A	1.25	1.25	1.5	1.5	0	0	0	0	—	—

place, the percentage representations would fluctuate randomly, as they do in early generations.

Examination of the top experiments reveals certain patterns. For example, most top experiments have common values for five of the experimental coordinates ($DNA = C$, $AA_0 = 1.25$, $AA_{30} = 4$, $T_1 = 120$, $T_2 = 30$). These values comprise a "building block" discovered by the learning algorithm, and they define a hyperspace of co-dimension five in the experimental space, a region of higher-than-average fitness. Figure 4 shows conditional fitness distributions (fitness conditioned on the number of experimental coordinates that match key values, defined as those values that are the same across the top ten experiments), which illustrate both the average higher fitness within the hyperspace, and the fact that substantial fitness increase is obtained only when three or more of these conditions are present simultaneously indicating that these elements are synergetic in the sense that they produce high fitness only by working together. Most of the conditions alone are seen to give a statistically significant increase in average fitness, but this increase is extremely small when compared to the gain when all conditions are satisfied together. Further analysis has shown that the most important key variable was the DNA type.

The evidence of significant synergy between experimental components in Figure 4 contradicts a conclusion drawn from a different in vitro protein synthesis experiment (Matsuura et al., 2009), where a coarse-grained statistical model using the Bahadur expansion truncated after the pairwise interaction terms could yield a value of the coefficient of determination $R^2 > 99\%$ for the correlations between predicted and observed responses (calculated on

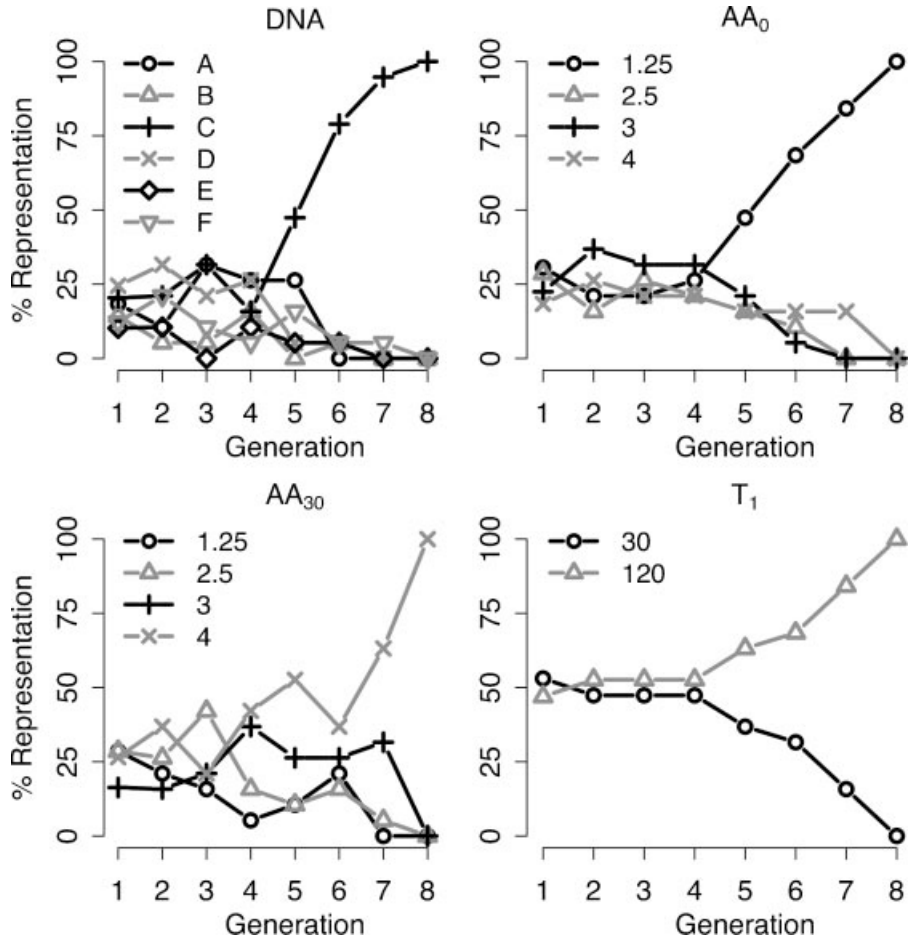

Figure 3. Evolutionary learning dynamics: representation of particular experimental coordinate values in the population of experiments, for each generation.

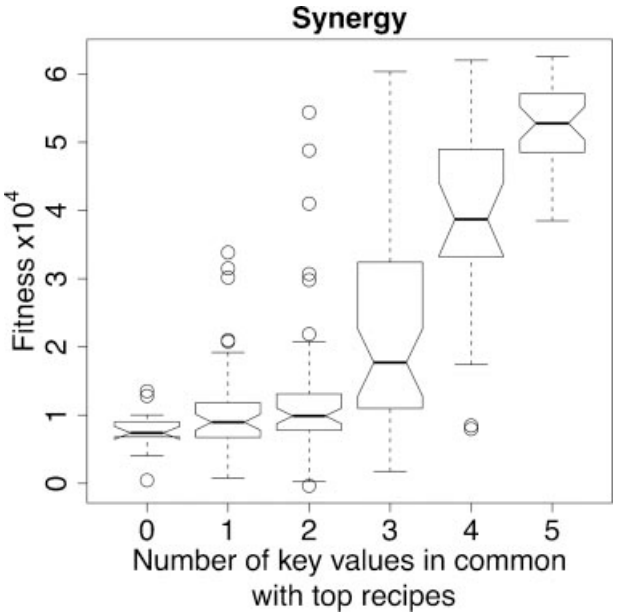

Figure 4. Illustration of synergy: conditional distributions of fitness over all experiments. Notches on each box show 95% confidence intervals around the median. Top experiments have key values for certain of the experimental variables ($DNA = C$, $AA_0 = 1.25$, $AA_{30} = 4$, $T_1 = 120$, $T_2 = 30$). The distributions shown are conditioned on the number of experimental coordinates that match the key values.

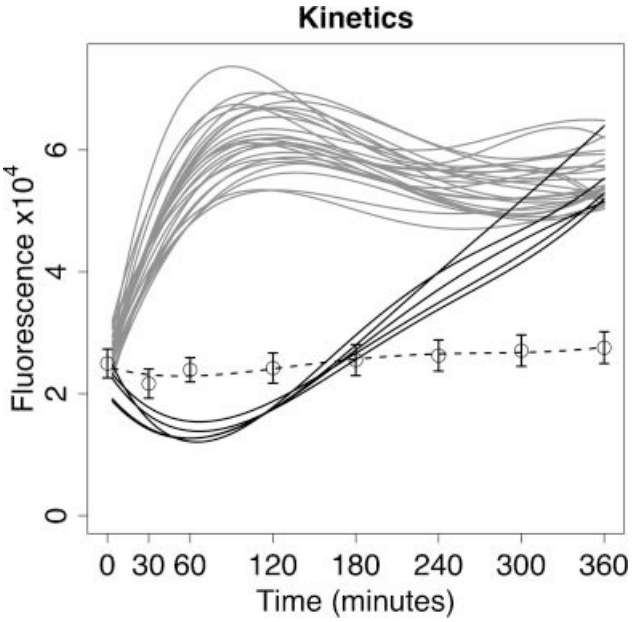

Figure 5. Two distinct classes of kinetics are observed in the top 30 experiments. The standard is shown with a dotted line, with error bars representing the standard deviation from all repeats. Each curve is a smoothed quartic fit of fluorescence measurements taken at the eight time points as superimposed on the standard line. Experiments with class one kinetics (plateau) are shown in gray, those with class two kinetics (increasing) are shown in black.

64 observations). A similar analysis of our data, using a polynomial model with the same terms as the above Bahadur expansion, indicates $R^2 \approx 90\%$ (calculated on 215 observations), and thus more strongly indicates the presence of higher-order interactions in our system. However, it is important to emphasize that the sample of the experimental space is too sparse to make strong statements regarding the order of the interactions, and in our case, the sample is strongly affected by the model-based evolutionary learning procedure used to explore the space.

Substantial insight into the structure of the fitness landscape is gained by observing the kinetics of each experiment, that is, the temporal variation of fluorescence that indicates product yield. These kinetics are illustrated in Figure 5, for the top 30 experiments ordered by final fluorescence. Remarkably, we observe two distinct classes of kinetics, both discovered by the predictive evolutionary algorithm, and both kinetic classes differ substantially from the standard, which gave a comparatively flat, low response. The first class has kinetics that reach a plateau by the end of the observations ($t = 360$ min), after reaching a maximum around $t \approx 120$ min. The second class of kinetics displays an initial decrease in yield, followed by a strong increase that continues to the end of the observations. The high final slope of the kinetics of this second class of experiments suggests that their maximal product yield (fitness) might have been substantially higher than any other experiments if our observation period had been longer. This also suggests that more selective pressure could be placed on the system to improve the yield even further. It should be noted that the fitness function did not rate class two kinetics with a high score, because the best class one kinetics had a maximum fluorescence value peak that exceeded the maximum value achieved by the class two kinetics, as shown in Figure 5.

The underlying mechanism that explains these two classes can be elucidated upon further experimentation. The eGFP fluorescence on which our fitness function is based is modulated by the efficiency of transcription, stability of the mRNA, efficiency of translation, and folding and stability of the protein. Each of these steps could be influenced by the parameters varied in our experimental system. Therefore it is not surprising that very different kinetic profiles can result as experimental variables change. The shape of the line for the first kinetic class (gray, Fig. 5) appears as a standard unregulated gene expression profile. However, the profile for the second kinetic class (black, Fig. 5) shows delayed expression, as if an inducer was added late to the system. No such inducer was added, so perhaps the initial condition resulted in suppression of protein expression which was later overcome by the introduction of high amounts of magnesium salt at $t = 120$ min (compare Tables III and IV). Also, it has been noted in at least one case that an initially high concentration of amino acids (exemplified in the second kinetic class) may lead to inhibition of protein production (Jefferson and Korner, 1969). We find, however, in Figure 3, that optimal yield is obtained with the lowest

Table IV. Top four experiments with class two kinetics, ordered by final value of fluorescence.

Rank	*DNA*	AA_0	AA_{30}	M_0	M_{30}	*PEP*	*Mg*	*G*	*NADH*	T_1	T_2
Top experiments with class two kinetics											
1	F	4	4	3	3	0	30	60	0	120	30
2	F	4	4	4.8	3.6	23	23	60	0	120	30
3	D	4	1.25	1.5	1.5	30	30	15	0	120	30
4	C	4	4	3.6	3.6	0	23	15	0	120	30
5	B	4	4	4.8	4.8	12	30	0	0.32	120	30

concentration of amino acids at the start of an experiment and the highest amount then added after 30 min. A priori, one would assume that maximizing the amount of amino acids added to the system would maximize protein output. Using our optimization approach, we were able to discover a counterintuitive preference for the least amount of amino acids at the beginning of the experiments in order to produce the maximum amount of eGFP production. This may support the case for inhibition of the initiation of protein expression by high concentrations of amino acids. A careful characterization of the protein expression system with the defined parameters discovered here should reveal the underlying causes of this variation.

Table IV shows experiments from the top five of the second kinetic class, ordered by final fluorescence. We see that these experiments lie outside the high yield hyperspace observed in the top experiments, sharing with the top experiments only two out of the five key values ($T_1 = 120$, $T_2 = 30$), and partially one of the other key values ($AA_{30} = 4$). These experiments also show strong representation of DNA values of types D and F, neither of which are seen in the top experiments, and the latter of which (type F) has spacer lengths that inhibit simultaneous docking of the RNA polymerase and mRNA.

Visualization of high-dimensional spaces is not possible directly, but some structure maybe be seen in Figure 6 shows a dendrogram representation that uses clustering. All experiments are represented as points in a Euclidean space, with their coordinates normalized so that all points lie within the unit cube. Then the closest points (using the Euclidean norm) are aggregated to form clusters of pairs, closest pairs are aggregated, and so on, till the aggregation includes the entire space. Top experiments from the first kinetic class are identified with circles, and those from the second class with crosses. Evidently, the experiments from the first kinetic class are more localized, and those from the second kinetic class are spread more throughout the space, that is, they occur in many clusters. This might indicate that the experiments in the second kinetic class have more potential for further optimization.

Regarding the DNA that was selected by the optimization process, most templates perform similarly but sequence C outperforms other templates significantly. This template is characterized by a long spacer between RNA polymerase promoter and RBS (L_1, 20 base pairs) and between the RBS and start codon of the encoding region (L_2, 7 base pairs).

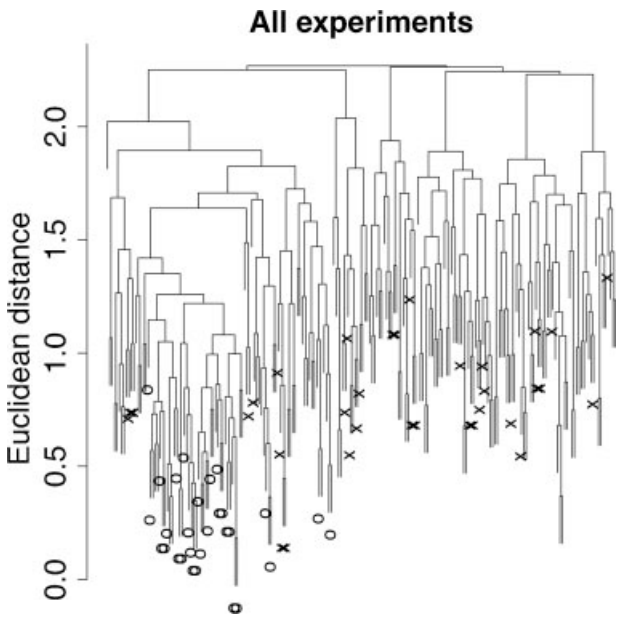

Figure 6. Dendrogram representation of all experiments showing clustering of the first kinetic class. The experimental space was transformed so that all experiments lie in a unit cube, and the Euclidean metric was used to form the distances, which were hierarchically clustered to build the dendrogram. Top experiments by value of fitness with class one kinetics are indicated with circles, and all experiments with class two kinetics are indicated with crosses.

The observation of enhanced performance for this sequence can be explained by the assumption that long spacers minimize the unfavorable interaction between RNA polymerase and ribosome minor subunits during the coupled transcription/translation process. Long spacers might therefore exert a positive synergistic effect increasing overall process kinetics and yield. This explanation is supported by the observation that DNA templates with short spacers (e.g., template F) display a slow initial kinetic with a prolonged lag phase (the black curves in Fig. 5). It is noteworthy that on average DNA template E—characterized by a stable hairpin secondary that masks the RBS—performs as poorly as the negative standard sequence F—characterized by suboptimal spacers—and significantly worse than the correspondent template with readily accessible RBS (template A). Some exceptions to this average behavior may also be observed; remarkably, the negative standard sequence F appears in the experiment with the strongest class two kinetics (cf. Table IV). This supports the hypothesis that secondary structure at the mRNA level may significantly affect overall performance as much as improper mutual location of regulatory elements on the DNA template. Although DNA template design seems to be a key parameter, absolute fitness and output reproducibility is deeply affected by other experimental parameters, further supporting our conclusion that optimal expression is ultimately described by synergic effects of several components.

The experiments were quite reproducible, as seen in many experimental repeats, both spatially (in different wells of the 384-well plate) and temporally (on different days). The noise level on the repeats is seen explicitly in the error bars around the standard, in Figure 5, and the noise level for other repeats was comparable.

Conclusion

We have shown that complex experiments such as protein synthesis may be optimized using an evolutionary learning technique that combines stochastic exploration with exploitation of structure discovered in the data by construction of a predictive model that can be sampled as a form of virtual experimentation. Examination of the kinetics of the highest-yield experiments reveals that two distinct classes of kinetics were discovered: one that reaches a maximum fluorescence and then plateaus, and another that shows an initial decrease in fluorescence followed by a strong increase throughout the observation period. The first class of kinetics was relatively localized in the space of experiments; the second was found in many separate regions of the space. The space of possible experiments was large, $\sim 1.5 \times 10^6$, and an increase in yield of over 300% was obtained by sampling only ~0.014% of the space. There is no guarantee that the best experimental results in the space have been discovered; indeed, we expect that further optimization is likely possible.

We believe that the optimized protocol found here for specifically eGFP synthesis may be suboptimal for other proteins and types of protein (i.e., hydrophobic and membrane-bound proteins). However, the optimization method (Evo-DoE) is general enough to optimize the synthesis of other proteins. In addition, we expect that if the constraints to encourage generalizability between proteins were relaxed (e.g., individual variation of each amino, or even the addition of non-canonical amino acids), further optimization of a specific protein of interest should be possible.

Evo-DoE may be compared to recent efforts to automate scientific research using machine learning combined with robotic technology (Waltz and Buchanan, 2009), namely, the automated analysis of large quantities of experimental data that is becoming available through genomics and proteomics (King et al., 2004, 2009), and the discovery of physical laws of mechanics (Schmidt and Lipson, 2009). Our motivation is different; it aims to improve a desired experimental result directly. Hypothesis generation for us is secondary; we do not build and test an explicit logical model relating system components. The only hypothesis we generate is a set of new experiments, proposed by a statistical model built from the data. In one way our goal is more modest that the previous work, because we do not aim to fully automate the scientific process and replace the human scientist. For us, the human scientist provides essential input in the design of the experimental space to be explored, and finally, in the interpretation of results. However, in another way our goal is more ambitious. Instead of achieving results merely comparable to human scientists, we achieve results that humans scientists find virtually impossible to obtain without our techniques.

Complex experiments such as protein synthesis tend to yield limited results when directed by systematic application of fundamental theory. We have shown how to direct experiments by an automated intelligent Edisonian process of trial and error. We expect that the increasing interest in complex experiments and the continued improvement in robotic automation of high-throughput experimentation will lead to widespread use of our Edisonian approach.

The authors acknowledge John McCaskill and Hans Ziok for helpful comments on the manuscript. The authors also acknowledge the European Center for Living Technology for workshops on related topics. NP and GG acknowledge helpful conversations with Irene Poli. FC, GG, MH, and NP acknowledge helpful conversations with Tomoaki Matsuura.

References

Abe M, Ohno S, Yokogawa T, Nakanishi T, Arisaka F, Hosoya T, Hiramatsu T, Suzuki M, Ogasawara T, Sawasaki T, Nishikawa K, Kitamura M, Hori H, Endo Y. 2007. Detection of structural changes in a cofactor binding protein by using a wheat germ cell-free protein synthesis system coupled with unnatural amino acid probing. Proteins 67(3): 643–652.

Anderson CW, Straus JW, Dudock BS. 1983. Preparation of a cell-free protein-synthesizing system from wheat germ. Methods Enzymol 101:635–644.

Breiman L. 1996. Bagging predictors. Machine Learn 24(2): 123–140.

Calhoun KA, Swartz JR. 2005. Energizing cell-free protein synthesis with glucose metabolism. Biotechnol Bioeng 90:606–613.

Calhoun KA, Swartz JR. 2007. Energy systems for ATP regeneration in cell-free protein synthesis reactions. Methods Mol Biol 375:3–17.

Caschera F, Gazzola G, Bedau MA, Bosch Moreno C, Buchanan A, et al. 2010. Automated discovery of novel drug formulations using predictive iterated high throughput experimentation. PLoS ONE 5(1): e8546.

Cawse JN, Gazzola G, Packard N. 2010. Efficient discovery and optimization of complex high-throughput experiments. Catal Today 159(2011): 55–63.

Chambers TJ, Nestorowicz A, Amberg SM, Rice CM. 1993. Mutagenesis of the yellow fever virus NS2B protein: Effects on proteolytic processing. NS2B-NS3 complex formation, and viral replication. J Virol 67:6797–6807.

Corma A, Serra JM, Serna P, Valero S, Argente E, Botti V. 2005. Optimisation of olefin epoxidation catalysts with the application of high-throughput and genetic algorithms assisted by artificial neural networks (soft-computing techniques). J Catal 229:513–524.

Cornish VW, Benson DR, Altenbach CA, Hideg K, Hubbell WL, Schultz PG. 1994. Site-specific incorporation of biophysical probes into proteins. Proc Natl Acad Sci USA 91(8): 2910–2914.

Delbecq S, Hadj-Kaddour K, Randazzo S, Kleuskens J, Schettersa T, Gorenf lota A, Précigouta. 2006. Hydrophobic moeties in recombinant proteins are crucial to generate efficient saponin-based vaccine against Apicomplexan Babesia divergens. Vaccine 24:613–621.

Forlin M, Poli I, De March D, Packard N, Gazzola G, Serra R. 2008. Evolutionary experiments for self-assembling amphiphilic systems. Chemom Intell Lab Syst 90:153–160.

Forstner M, Leder L, Mayr LM. 2007. Optimization of protein expression systems for modern drug discovery. Expert Rev Proteomics 4:67–78.

Goshima N, Kawamura Y, Fukumoto A, Miura A, Honma R, Satoh R, Wakamatsu A, Yamamoto J-i, Kimura K, Nishikawa T, Andoh T, Iida Y, Ishikawa K, Ito E, Kagawa N, Kaminaga C, Kanehori K-i, Kawakami B, Kenmochi K, Kimura R, Kobayashi M, Kuroita T, Kuwayama H, Maruyama Y, Matsuo K, Minami K, Mitsubori M, Mori M, Morishita R, Murase A, Nishikawa A, Nishikawa S, Okamoto T, Sakagami N, Sakamoto Y, Sasaki Y, Seki T, Sono S, Sugiyama A, Sumiya T, Takayama T, Takayama Y, Takeda H, Togashi T, Yahata K, Yamada H, Yanagisawa Y, Endo Y, Imamoto F, Kisu Y, Tanaka S, Isogai T, Imai J, Watanabe S, Nomura N. 2008. Human protein factory for converting the transcriptome into an in vitro-expressed proteome. Nat Methods 5:1011–1017.

He M, Wang MW. 2007. Arraying proteins by cell-free synthesis. Biomol Eng 24:4375–4380.

Iffland A, Kohls D, Low S, Luan J, Zhang Y, Kothe M, Cao Q, Kamath AV, Ding YH, Ellenberger T. 2005. Structural determinants for inhibitor specificity and selectivity in PDE2A using the wheat germ in vitro translation system. Biochemistry 44:8312–8325. Biotechnol Bioeng 104: 1189–1196.

Jefferson LS, Korner A. 1969. Influence of amino acid supply on ribosomes and protein synthesis of perfused rat liver. Biochem J 111:703–712.

Jiang ZG, Liu Y, Hussain MM, Atkinson D, McKnight CJ. 2008. Reconstituting initial events during the assembly of apolipoprotein B-containing lipoproteins in a cell-free system. J Mol Biol 383:1181–1194.

Jones DR, Schonlau M, Welch WJ. 1998. Efficient global optimization of expensive black-box functions. J Global Optim 13:455–492.

Katzen F, Chang G, Kudlicki W. 2005. The past, present and future of cell-free protein synthesis. Trends Biotechnol 23:150–156.

Kim DM, Swartz JR. 1999. Prolonging cell-free protein synthesis with a novel ATP regeneration system. Biotechnol Bioeng 66:180–188.

Kim D-M, Swartz JR. 2000. Prolonging cell-free protein synthesis by selective reagent additions. Biotechnol Prog 16:385–390.

Kim DM, Swartz JR. 2001. Regeneration of adenosine triphosphate from glycolytic intermediates for cell-free protein synthesis. Biotechnol Bioeng 74:309–316.

Kim DM, Kigawa T, Choi CY, Yokoyama SA. 1996. Highly efficient cell-free protein synthesis system from *E. coli*. Eur J Biochem 239:881–886.

Kim TW, Kim DM, Choi CY. 2006. Rapid production of milligram quantities of proteins in a batch cell-free protein synthesis system. J Biotechnol 124:373–380.

Kim T-W, Keum J-W, Oh I-S, Choi C-Y, Kim H-C, Kim D-M. 2007a. An economical and highly productive cell-free protein synthesis system utilizing fructose-1,6-bisphosphate as an energy source. J Biotechnol 130:389–393. Mental data. Science 324: 81–85.

Kim T-W, Oh I-S, Keum J-W, Kwon Y-C, Byun J-Y, et al. 2007b. Prolonged cell-free protein synthesis using dual energy sources: Combined use of creatine phosphate and glucose for the efficient supply of ATP and retarded accumulation of phosphate. Biotechnol Bioeng 97:1510–1515.

King RD, Whelan KE, Jones FM, Reiser PGK, Bryant CH, Muggleton SH, Kell DB, Oliver SG. 2004. Functional genomic hypothesis generation and experimentation by a robot scientist. Nature 427:247–252.

King RD, Rowland J, Oliver SG, Young M, Aubrey W, Byrne E, Liakata M, Markham M, Pir P, Soldatova LN, Sparkes A, Whelan KE, Clare A. 2009. The automation of science. Science 324:85–88.

Klarmann GJ, Eisenhauer BM, Zhang Y, Gotte M, Pata JD, Chatterjee DK, Hecht SM, Le Grice SF. 2007. Investigating the ''steric gate'' of human immunodeficiency virus type 1 (HIV-1) reverse transcriptase by targeted insertion of unnatural amino acids. Biochemistry 46:2118–2126.

Lesley SA, Brow MA, Burgess RR. 1991. Use of in vitro protein synthesis from polymerase chain reaction-generated templates to study interaction of *Escherichia coli* transcription factors with core RNA polymerase and for epitope mapping of monoclonal antibodies. J Biol Chem 266:2632–2638.

Liguori L, Marques B, Lenormand JL. 2008. A bacterial cell-free expression system to produce membrane proteins and proteoliposomes: From cDNA to functional assay. Curr Protoc Protein Sci 54:5.22.1–5.22.30.

Liping D, Rong G, Forster AC. 2009. Engineering multigene expression in vitro and in vivo with small terminators for T7 RNA polymerase. Biotechnol Bioeng 104:1189–1196.

Matsuura T, Kazuta Y, Aita T, Adachi J, Yomo T. 2009. Quantifying epistatic interactions among the components constituting the protein translation system. Mol Syst Biol 5:1–10.

Nirenberg MW, Matthaei JH. 1961. The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. Proc Natl Acad Sci USA 47:1588–1602.

Osada E, Shimizu Y, Akbar BK, Kanamori T, Ueda T. 2009. Epitope mapping using ribosome display in a reconstituted cell-free protein synthesis system. J Biochem 145:693–700.

Oyama R, Takashima H, Yonezawa M, Doi N, Miyamoto-Sato E, Kinjo M, Yanagawa H. 2006. Protein-protein interaction analysis by C-terminally specific fluorescence labeling and fluorescence cross-correlation spectroscopy. Nucleic Acids Res 34(14): e102.

Ozawa K, Headlam MJ, Schaeffer PM, Henderson BR, Dixon NE, Otting G. 2004. Optimization of an *Escherichia coli* system for cell-free synthesis of selectively N-labelled proteins for rapid analysis by NMR spectroscopy. Eur J Biochem 271:4084–4093.

Pelham HRB, Jackson RJ. 1976. An efficient mRNA-dependent translation system from reticulocyte lysates. Eur J Biochem 67:247–256.

Sasena MJ, Parkinson M, Reed MP, Papalambros PY, Goovaerts P. 2005. Improving an ergonomics testing procedure via approximation-based adaptive experimental design. J Mech Des 127:1006–1013.

Schmidt M, Lipson H. 2009. Distilling free-form natural laws from experimental data. Science 324:81–85.

Sitaraman K, Chatterjee DK. 2009. High-throughput protein expression using cell-free system. Methods Mol Biol 498:229–244.

Spirin AS, Baranov VI, Ryabova LA, Ovodov SY, Alakhov YB. 1998. A continuous cell-free translation system capable of producing polypeptides in high yield. Science 242:1162–1164.

Stein ML. 1999. Interpolation of spatial data: Some theory for kriging. New York: Springer.

Studier FW, Rosenberg AH, Dunn JJ, Dubendorff JW. 1990. Use of T7 RNA polymerase to direct expression of cloned genes. Methods Enzymol 185:60–89.

Theis M, Gazzola G, Forlin M, Poli I, Hanczyc M, Bedau M, 2006. Optimal formulation of complex chemical systems with a genetic algorithm. In: Jost J, Reed-Tsochas F, Schuster P, editors. ECCS06 Online Proceedings. Available at: http://sbs-xnet.sbs.ox.ac.uk/complexity/complexity_PDFs/ECCS06/Conference_Proceedings/PDF/p193.pdf.

Underwood KA, Swartz JR, Puglisi JD. 2005. Quantitative polysome analysis identifies limitations in bacterial cell-free protein synthesis. Biotechnol Bioeng 91(4): 425–435.

Waltz D, Buchanan BG. 2009. Automating science. Science 324:43–44.

Wang X, Liu J, Zheng Y, Li J, Wang H, Zhou Y, Qi M, Yu H, Tang W, Zhao WM. 2008. An optimized yeast cell-free system: Sufficient for translation of human papillomavirus 58 L1 mRNA and assembly of virus-like particles. J Biosci Bioeng 106:8–15.

Woodrow KA, Airen IO, Swartz JR. 2006. Rapid expression of functional genomic libraries. J Proteome Res 5:3288–3300.

Zubay G. 1973. In vitro synthesis of protein in microbial systems. Annu Rev Genet 7:267–287.