THE FACULTY OF SCIENCES

MASTER IN BIOTECHNOLOGY

Experimental combinatorial optimization of phycobiliproteins׳ expression in *E. coli*

Chen Erlich

UNDER THE SUPERVISION OF Dr. Dror Noy and Dr. Ofer Shir
Migal- Galilee Research Institute

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR MASTER DEGREE IN BIOTECHNOLOGY

February 2019

Acknowledgements

I would first like to thank my thesis advisors Dr. Dror Noy and Dr. Ofer Shir of the Migal- Galilee Research Institute. Their door office was always open whenever I ran into a trouble spot or had a question about my research or writing.

I would also like to thank Mr. Assaf Israeli for the enormous help with analyzing the data and creating great figures that explains excellent the results.

I would also like to acknowledge Dr. Svetlana Yom-Din, Laboratory manager and mental supporter and all the colleagues from Migal.

Finally, I must express my very profound gratitude to my parents and to my fiancé for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them.

Thank you!

Chen Erlich

Table of Contents

Abstract

High throughput screening (HTS) assays target biological systems of high complexity that require searching through a wide range of parameters. In such cases, optimization routines that search over a single parameter at a time or specific subsets of parameters are ineffective. Thus, designing assays optimized for accuracy, speed, efficiency, high sensitivity, and low reagent consumption is critical for successful HTS campaigns. A novel design approach is to consider screening as a *combinatorial optimization* problem and apply an algorithmically-guided search over the entire array of parameters. In most biological systems, the dependence of the optimization target on the experimental parameters cannot be determined or simulated explicitly, which requires evaluating the target by actual measurements. This class of optimization problems is referred to as *experimental optimization*. In this work, we implemented experimental optimization for optimizing the heterologous expression of Allophycocyanin A (ApcA), a cyanobacterial phycobiliprotein, in *E. coli*. Phycobiliproteins are water-soluble proteins that incorporate linear tetrapyrrole pigment molecules (chromophores), called phycobilins by covalently binding to specific cysteine residues. In cyanobacteria and red algae, they self-assemble with other chromophore-free proteins to form enormous protein complexes, known as phycobilisomes. The organization of the phycobiliprotein subunits into an elaborate network of protein-pigment complexes provides the photosynthetic organisms with highly robust and dynamic light-harvesting system that effectively absorbs light and transfers its energy. Beside their role as building blocks of photosynthetic light harvesting complexes, there has been much interest in phycobiliproteins as spectroscopic markers in medical and biotechnological applications.

ApcA binds a single phycocyanobilin chromophore, and is part of the red-most absorbing subunit in the core of the phycobilisome. Its heterologous overexpression system includes four genes: one for expressing the ApcA apo-protein, two genes that are required for phycocyanobilin biosynthesis and the fourth gene for expressing a lyase catalyzing the covalent linking of phycocyanobilin to the protein. Such a system is well suited for experimental optimization because 1) its efficiency depends on many parameters, some of which are easily controllable external parameters, while others are more difficult to control internal parameters, 2) the optical spectroscopic signatures (emission and absorption spectra) of the desired product can be measured with low background, and a high signal-to-noise ratio, 3) the yield of holo-phycobiliprotein can be mathematically formulated as an objective function, and 4) experimental protocols are well established, relatively inexpensive and not highly time consuming.

In this study, we implemented experimental combinatorial optimization in the aim of increasing the production yield of recombinant ApcA-phycocyanobilin holo-protein. To this end, we defined an objective function based on optical measurements of the protein samples, and an array of ten external parameters (*decision variables*) that can affect the holo-protein production. The definitions and values were coded in an algorithm which initialized a random population of combinations of experimental parameters and iteratively applied variation operators to construct an offspring population, out of which the consecutive parental population was selected based on the objective function value obtained for each candidate combination. We established an experimental protocol that received the combination of parameters from the computational algorithm, implemented them in a protein expression and purification system, and measured the resulting objective function for each combination. The result was then fed back to the optimization algorithm for generating the next generation of candidate combinations. Altogether, an initial generation comprising 48 candidate combinations, and five more generation each comprising 24 candidate combinations were tested. We observed an improvement in the objective function values from one generation to another, but the maximal yield was lower than expected. The optimization method, which relies on objective values' ranking, evidently increased the total protein expression yield while not significantly increasing the ratio of holo-protein to total protein. Detailed analysis of the entire campaign revealed that it is likely due to inaccurate definition of the objective function, which was formulated according to the expert's intuition. Better computational tools may be required to devise refined mathematical formulations that properly reflect the specific protein expression targets. This notwithstanding, the current campaign provided insights into the biological system. We discovered surprising and even counter-intuitive parameters that could improve protein production, such as avoiding the use of IPTG under the control of the T7 promotor. This specific insight gives us a hint about variations at the molecular biology level that correspond to yield improvements obtained by the optimization procedure (e.g., using a promotor allowing a tighter regulation of protein expression rather than the T7 promotor).

Although the algorithm did not maximize the protein production as desired, we gained many insights into both the mathematical and biological aspects of the system which will assist in planning and defining more accurate objective functions and decision variables in future experimental optimization campaigns.

Keywords: objective function, heterologous expression system, ApcA, holo-protein

Abrreviations

Γ-ALA- Γ-Aminolevulinic Acid

Å- Angstrom

APC- Allophycocyanin; ApcA- Allopchycocyanin Alpha (subunit)

DNA- Deoxyribonucleic Acid

DNB- Diluted Nutrient Broth

E. coli- Escherichia Coli

HO- Heme Oxygenase

Hr- Hour(s)

HTS- High Throughput Screening

IPTG- Isopropyl β-D-thiogalactopyranoside

kHZ- Kilo Hertz

LB- Luria Broth

M- Molar

ml- milliliter

mM- millimolar

nm- nanometer

O.D- Optical Density

PC- Phycocyanin

PCB- Phycocyanobilin

PcyA- Phycocyanobilin: Ferredoxin oxidoreductase

PE- Phycoerythrin

PEB- Phycoerythrobilin

PEC- Phycoerythrocyanin

rpm- revolutions per minute

TB- Teriffic Broth

UV- Ultra-Violet

List of Tables

List of Figures

1. Introduction

1.1 High Throughput Screening

Most assays routinely used in basic biochemistry, cell and molecular biology, human genetics, and other fields of basic research are not suitable for industrial-scale screening. This situation has led to the development of a specialized discipline devoted to the design of assays that are optimized for speed, efficiency, signal detection and low reagent consumption known as *high throughput screening (HTS)***.** Assays for HTS not only require small sample volume, high throughput and robustness, but also adequate sensitivity, reproducibility and accuracy in order to discriminate among a very large number of compounds or molecules that span a wide range of activities.

In most cases, the systems that are targeted by HTS are highly complex due to their inherent dependency upon a vast array of parameters that are of continuous, integral and categorical nature. For example, isolating antibodies against impure proteins and complex antigens, where several rounds of phage display often fail [1], or discovery of drug candidates for metabolism and pharmacokinetic characteristics [2]. This dependency renders tuning attempts by means of trial-and-error infeasible.

Using automation to systematically search over a single parameter at a time, or to analyze specific subsets of parameters, is more accurate since it explores the search-space. But, it is highly ineffective and often times mathematically-deficient for obtaining the broad picture. One way of making HTS more effective is by considering screening as a *combinatorial optimization* problem,[3,4] and applying an algorithmically-guided search over the entire array of parameters in order to identify the parameters with the best attainable target. Such search algorithms have been widely applied in the computer sciences to global optimization of complex models where the objective function either possesses an explicit expression or can be represented by a computer-based model. In the biological sciences, it is usually not possible to compute the objective function because rigorous modeling is infeasible, no simulation can be compiled, or the existing approximations are not sufficiently precise. In these cases, the objective function values are obtained by actual measurements of the experimental system in the laboratory. This class of optimization problems is referred to as *experimental optimization*[5] (the evaluation of candidate solutions is done by conducting a physical experiment) and it lies in the focal point of this work. Our primary goal is to implement this algorithmic approach in the optimization of a system for production of bio-molecules.

To be amenable to experimental optimization, the production system must satisfy the following requirements:

- ✓ A well-defined objective function – a reliable quantitative measure to a given process' efficiency, which reflects the scientist's target value.
- ✓ Control of the decision variables (i.e., the tuned parameters) at high level of accuracy.
- ✓ Tolerable statistical noise level of the measured values.
- ✓ Reasonable cost and time to execution of any measure/assay (experimental trial).

Figure 1 provides a scheme of a typical single iteration (step) in a proposed optimization run of a screening procedure targeting protein production yield. The input variables for each candidate solution are prescribed by the algorithm, whereas the output (feedback) is provided by a specific assay of protein production - altogether closing a feedback loop.

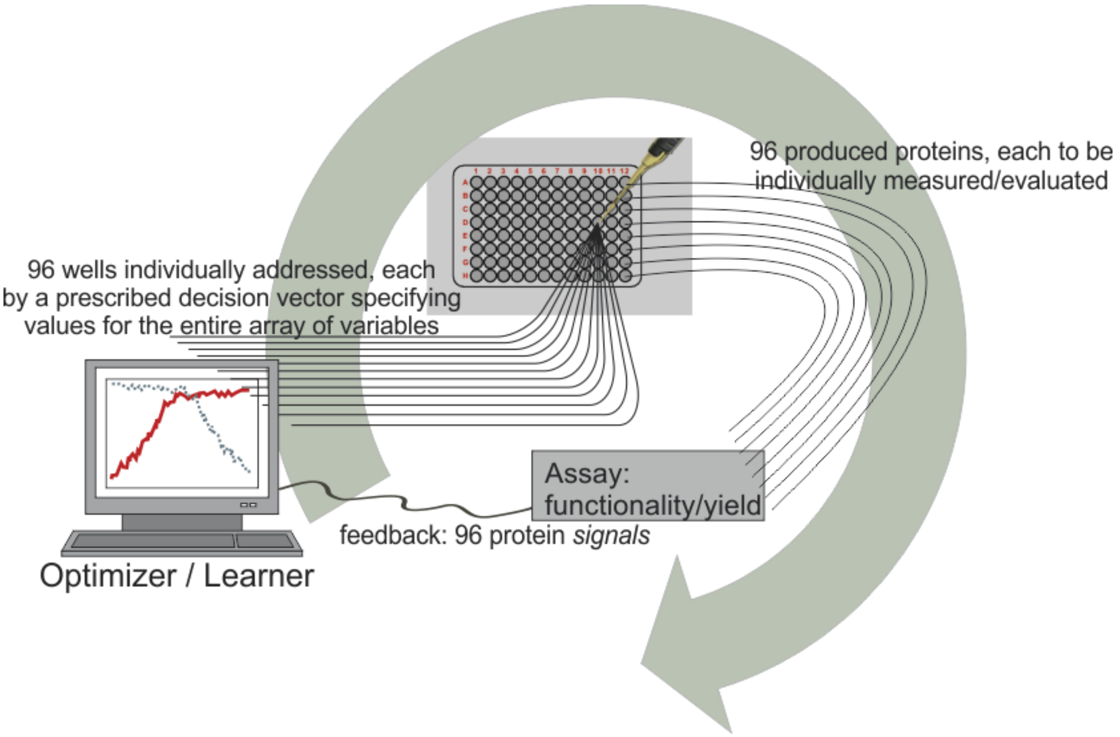

96 produced proteins, each to be individually measured/evaluated
1 2 3 4 5 6 7 8 9 10 11 12
A B C D E F G H
96 wells individually addressed, each by a prescribed decision vector specifying values for the entire array of variables
Assay: functionality/yield
feedback: 96 protein *signals*
Optimizer / Learner

Figure 1. A scheme of a typical single iteration in experimental combinatorial optimization. The 96-well plate represents an experiment comprising 96 samples. The screening of the plate gives the target values of the tested samples, which are then communicated to the optimization algorithm. The algorithm then devises the next 96 combinations, by deducing from failures and successes, which are likely to progress the search in the most effective manner.

This closed-loop approach was successfully applied in various studies of evolvable hardware and evolutionary robotics. A few studies applied evolutionary algorithms in microbiology, genomics, immunology and biochemistry contexts. One of these used a search algorithm to locate a drug combination in experiments identifying combinations of three doses of up to six drugs for selective killing of human cancer cells. Search algorithms resulted in a highly significant enrichment of selective combinations, when compared with a random search [6].

In this work, we chose the heterologous expression system of the phycobiliprotein allophycocyanin A (ApcA) in *E. coli* as the case study for the application of experimental optimization. The system is used in our laboratory in the study of phycobiliproteins assembly. Its efficiency depends on many factors, some of which are external, easily controllable parameters, such as growth temperature, induction temperature, and growth medium, while others are internal and more difficult to control, like the type of promotor and its sensitivity, or the type of the hosting cell and its ability to express non-native genes and proteins. Altogether, the requirements for applying combinatorial experimental optimization methods on the system are fulfilled as follows:

- ✓ The yield of holo-phycobiliprotein is a well-defined objective function.
- ✓ Various controllable external parameters that influence expression yields qualify as decision variables.
- ✓ The optical spectroscopic signature (emission and absorption spectra) of the desired product can be measured with low background, and a high signal-to-noise ratio.
- ✓ Experimental protocols are well established, relatively inexpensive, and not highly time consuming.

Thus, ApcA production in high yield by heterologous expression in *E. coli* is an appealing target for experimental optimization.

1.2 Structure, function and assembly of phycobilisomes

Phycobilisomes are elaborate light-harvesting systems found in cyanobacteria and red algae. Unlike other photosynthetic light-harvesting systems, the phycobilisomes are not an integral part of the photosynthetic membrane (thylakoid), but are connected to its cytoplasmatic side. These enormous protein complexes are responsible for the effective absorption of light, and the transfer of energy by organizing a network of protein-pigment subunits. The light energy absorbed by the peripheral subunits is transferred through the complex to its membrane-bound core. From there, it is transferred to the transmembrane photosystem Ⅱ where it is used to drive the initial photochemical processes that end up in water-splitting, charge separation and buildup of electrochemical potential across the thylakoids [7](Fig. 2). The phycobilisome is composed of phycobiliprotein subunits. These are water soluble proteins that are divided into two major types: the first incorporates linear tetrapyrrole pigment molecules (chromophores), called phycobilins, which are usually bound covalently to the protein through a specific amino acid residue (mostly cysteine) [8], and the second type, which usually lacks chromophores, is used for specific arrangement, connection, and organization of the different subunits within the phycobilisome complex. The chromophore-bearing subunits form heterodimers (AB), which further assemble into a ring-shaped trimer $(AB)_3$. Altogether, there are several types of

(AB)$_3$ heterodimeric rings that vary in their absorption and fluorescence properties, and the numbers and types of chromophores in each of their A or B subunits. The red-most absorbing types are allophycocyanins (APCs) that bind six chromophores per ring. The other types, namely, phycocyanins (PCs), phycoerythrocyanins (PECs), and phycoerythrins (PEs) absorb and fluoresce more to the blue, and contain nine, fifteen, or eighteen pigments per ring. Some of the linker subunits fit at the center of the trimer rings, assembling them into cylinders of different lengths, while others connect the different cylinders to each other and to the photosynthetic membrane.

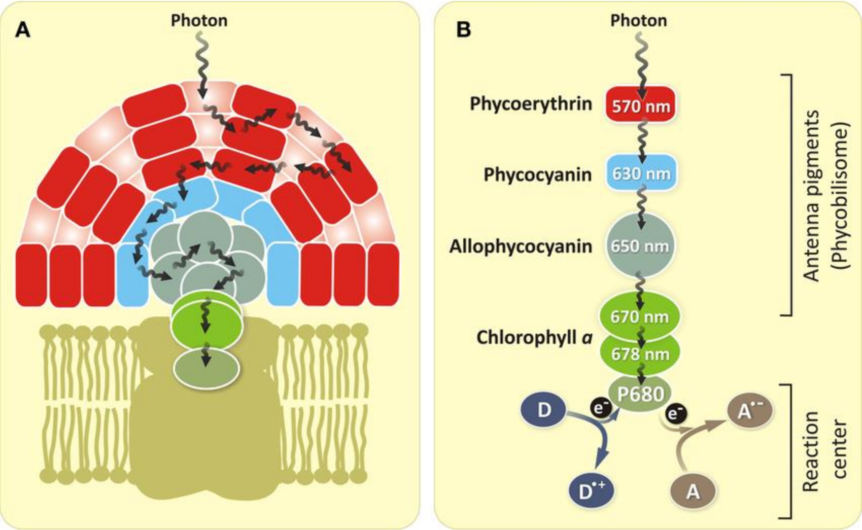


Figure 2. Structural organization of the antenna system of photosystem Ⅱ for red algae and cyanobacteria (A) and energy transfer steps including charge separation (photochemical reaction) at the PS Ⅱ RC (Reaction Center) (B) for cyanobacteria. (Adapted from: Govindjee; Shevela, D. Adventures with Cyanobacteria: A Personal Perspective. Front. Plant Sci. 2011, 2, 1-17).

The chromophores of the phycobilisomes are open-chain tetrapyrrole pigments known as phycobilins. These are derived from bilins, which are products of heme metabolism. There are several types of phycobilins that differ in the degree of conjugation of the tetrapyrrole π-system, which affects their light absorption and emission spectra. The most abundant phycobilins are phycocyanobilin (PCB) and phycoerythrobilin (PEB). Phycobilin biosynthesis follows the final step of heme biosynthesis, that is the insertion of iron into protoporphyrin IX catalyzed by the enzyme ferrochelatase. The first and common step to all phycobilin biosynthesis is the conversion of heme to biliverdin IXa by heme oxygenase (HO). This is followed by specific reduction at one or two out of three different positions by ferredoxin-dependent bilin reductases to yield the different phycobilin derivatives. PCB is generated by a single reductase, namely, phycocyanobilin:ferredoxin oxidoreductase (PcyA), which catalyses two reductions - at ring A and the 18-vinyl group (Fig. 3) (Pessaraki, et. Al 2005).

The final step in phycobiliprotein assembly is the post-translational modification step known as chromophorylation. In this process, the phycobilin chromophores are attached to specific cysteine sites in the apo-protein via thio-ether bonds. The reaction is catalyzed by specific lyases, although it may occur spontaneously with low accuracy and low yields. The lyase CpcS connects PCB to its specific binding sites in the A and B subunits of APC [8]. This results in a dramatic change of the light absorption and emission properties of the chromophores. Free bilins absorb light poorly and their excited states are very short-lived, leading to the dissipation of excitation energy as heat. In contrast, incorporation into specific protein environment turns these molecules into excellent photoreceptors: their light absorption increases by an order of magnitude, and their excited-state lifetime increases by four orders of magnitude [9].

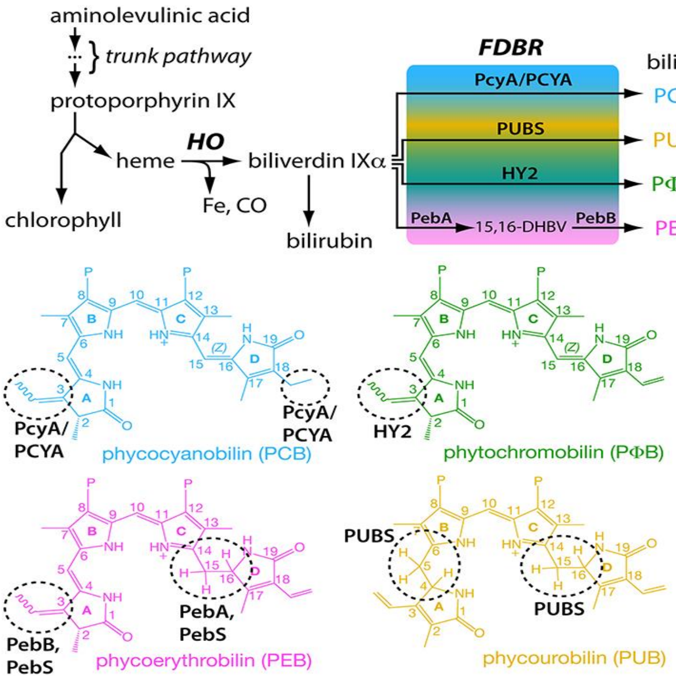


Figure 3. Biosynthesis of bilins, a common tetrapyrrole pathway increasing both heme and chlorophyll (top). Breakdown of heme proceeds via action of a heme oxygenase (HO) and a ferredoxin-dependent bilin reductase (FDBR). Different FDBRs carry out the same reaction (reduction) on different parts of the tetrapyrrole (bottom). This difference in regiospecificities allows production of a range of bilins with different spectral properties from a single precursor. (Adapted from: Rockwell, N. C., Lagarias, J. C., and Bhattacharya, D., Primary endosymbiosis and the evolution of light and oxygen sensing in photosynthetic eukaryotes, Frontiers in Ecology and Evolution, 2014, 2).

The supramolecular architecture, and subunit composition of the phycobilisomes is highly variable and depends on the specific organism, as well as on environmental conditions such as light quality and intensity, temperature, and available nutrients. All the phycobilisomes contain a core, comprising up to five short cylinders and made up of stacks of 2-3 APC rings. In some phycobilisomes, there are rods connected to the core, which form a radial section (Fig. 1). These rods are composed of stacked PC, PE, and PEC subunits. Advances in single-particle cryo-electron microscopy have recently enabled solving the structure of a whole phycobilisome from red algae at a resolution 3.5Å, thereby revealing the atomic details of a 16.8-megadalton phycobilisome complex comprising 862 protein subunits and 2048 chromophores (Fig. 3). This is a significant breakthrough in understanding phycobilisomes structure and assembly, yet, it should be noted that phycobilisome from other organisms may have different architectures.

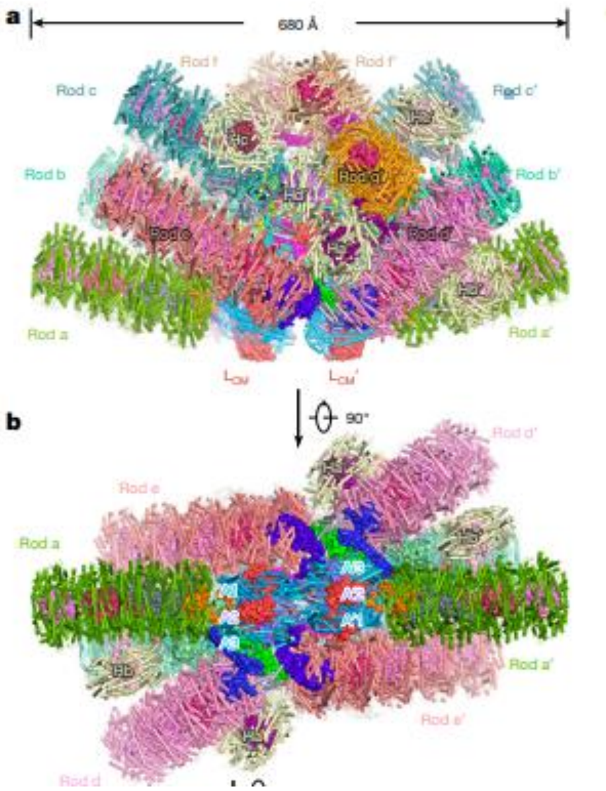


Figure 4. a-b: Structure of the phycobilisome from two perpendicular views. Phycobiliproteins are shown in cartoon representation, whereas linker proteins are shown as surface representation (Adapted from: Zhang J., Ma J., Liu D., Qin S., Sun S., Zhao J., and Sui S.F., Structure of phycobilisome from the red alga *Griffithisia pacifica*, Nature, Vol. 551, 57-84, November 2017)**.**

1.3 Commercial and medical applications of phycobiliproteins

In addition to their role as building blocks of photosynthetic light harvesting complexes, there has been much interest in phycobiliproteins as spectroscopic markers in medical and biotechnological applications. Flow cytofluorimetry, fluorescence microscopy and other fluorescence techniques have an important role in cell biology research. Phycobiliprotein conjugates with biologically active molecules can specifically bind to beads containing covalently attached target molecules and render them highly fluorescent. The main advantage of using such conjugates is the phycobiliprotein's specific light emission in the green and red regions of the optical spectrum, where background fluorescence is much lower than in the shorter near-UV and blue wavelengths [10].

Phycobiliproteins are non-carcinogenic and non-toxic, which makes them suitable for food and cosmetics applications. Phycocyanin is used as a natural colorant in food items, such as chewing gum, ice cream, dairy products, soft drinks (e.g. Pepsi® blue), soft candies and jellies as well as in cosmetics, such as lipsticks, eyeliners and eye shadows [11–13]. Phycobiliproteins, particulary phycocyanin, have a variety of potential pharmaceutical activities, such as antioxidant, anticancer, neuroprotective, anti-inflammatory, hepatoprotcetive, and hypocholoesterolemic [14–19]. Since most of the synthetic antioxidants currently in use are known to have side effects such as carcinogenesis and liver damage, antioxidants from natural sources are gaining impact as safe and effective alternatives to synthetic antioxidants. The antioxidant and radical scavenging activities of phycocyanin from different cyanobacteria are well documented [20–23].

1.4 Genetic systems for heterologous expression of phycobiliproteins

Heterologous expression of phycobiliproteins in *E. coli* is a powerful tool for rigorously studying phycobiliproteins structure and assembly, which is necessary for understanding their structure-function relationships within the photosynthetic apparatus. In addition, heterologous phycobiliproteins expression systems may open the way for exploiting the excellent biomarker properties of phycobiliproteins by using them as *in vivo* fluorescent protein probes [24]. Genetic systems for phycobiliprotein expression in *E. coli* have been developed by introducing genes for biosynthesis of phycobilins and their specific covalent binding to the apoprotein into the bacteria [24]. An example for such a system is the one developed by the Zhao laboratory for heterologous expression of APCs, which required introducing four foreign genes into the bacteria including:

a. A gene for expressing the specific APC apo-protein
b. A heme oxygenase (ho1) gene for converting protoheme to biliverdin molecule
c. A Phycocyanobilin:Ferredoxin Oxidoreductase (pcyA) gene for reducing the biliverdin at two sites to form PCB
d. A lyase gene for specific covalent linking of the chromophore to the protein.

The heterologous system was used in previous studies in order to explore questions in the enzymology and chemistry of phycobiliproteins synthesis. Zhao et. Al 2007 showed that the CpeS1 lyase has a broad substrate specificity for chromophorylating the Cys-84 site in almost all groups of cyanobacterial phycobiliproteins, which makes it a nearly universal lyase for the chromophorylation of apo-phycobiliprotein subunits [8]. More recently, the system was used to study the molecular basis and mechanism of red-shifting biliproteins spectra in certain cyanobacteria that contain red-shifted chlorophylls. It was found that only one allophycocyanin subunit is red-shifted, and that the shift is due to non-covalent binding of the phycocyanobilin chromophore [25]. In another study, the system was used to recombinantly express and assemble a water-soluble analogue of the terminal phycobilisome emitter, L_{CM}(ApcE). This was achieved by omitting a highly hydrophobic domain from the native protein sequence. This change did not affect autocatalytic binding of the chromophore, and the spectroscopic characteristics of the full-length protein and the result was modeling a subunit of cyanobacterial phycobilisome subunit. This water-soluble version of ApcE, was crystalized and its structure was solved revealing a unique conformation of the phycocyanobilin chromophore within the protein that is the cause for its spectral red-shift [26].

Another application of the heterologous system is to create semi-artificial, and hybrid phycobiliproteins. A recent example is the fusion of the chromophore-binding domain of ApcE to HP7, a de-novo designed heme, and chlorophyll-binding protein thereby assembling a multiple pigment complex analogous to natural light-harvesting complexes [27]. In a different study, various phycobiliprotein analogues were generated in situ from a single, multifunctional gene and endogenous heme. The construct consists of two persistently red-fluorescent biliproteins based on allophycocyanins and photochromic phycobiliproteins with special properties derived from unique cyanobacteriochrome [28].

A significant disadvantage of the system is the low yield of phycobiliprotein assembly, which usually does not exceed 10% of the total phycobiliprotein. Increasing the yield of holo-phycobiliprotein in heterologous expression systems is of great value because it will increase the ability to use phycobiliprotein's photophysical, spectral and fluorescence characteristics in many science fields such as cell biology, biochemistry and pharmacology.

2. Project objectives

2.1 Hypothesis

Applying a systematic search, as in experimental combinatorial optimization, will significantly improve protocols of biological production methods. Particularly, with proper experimental design, it should be possible to increase protein production yields in recombinant expression systems.

2.2 Specific Objectives

To test our hypothesis we implemented experimental combinatorial optimization heuristics for maximizing the heterologous expression of ApcA in *E. coli*. To this end, we set the following specific objectives:

- ✓ Construction of a heterologous expression system of ApcA in *E. coli* bacteria, which allows for scanning of multiple different growth and expression conditions.
- ✓ Establishing protocols for quickly cleaning and isolating the desired protein that can be used to analyze many samples at once.
- ✓ Defining a quantitative measure for the expression quality to be used as the objective function value within the optimization process.

- ✓ Implementing an optimization algorithm that aims at maximizing the objective function value by iteratively devising candidates of growth and expression parameters according to measured objective function values.

3. Materials and methods

3.1 Application of optimization procedure to the experimental system

The calculations and algorithmic-steps were carried out based on a heuristic implemented in Dr. Ofer Shir's group. This heuristic belongs to the so-called family of *Evolution Strategies*,[29] since it mimics the basic principles of natural evolution, where a large number of successive, slight modifications lead to the formation of highly specialized and elaborate systems. In short, it initializes a random population of combinations of experimental parameters, and iteratively applies variation operators to construct an offspring population, out of which the consecutive parental population is selected. The sole selection criterion is the objective function value of each candidate combination.

Applying the heuristic to a specific problem requires formulating an objective function to be used by the optimization process. Our goal was to obtain the maximal amount of ApcA holoproteins with the minimal amount of apo-protein. This required formulating an objective function that reflects our goal, or target value in a mathematical form that can be used by the optimization process. In our case, the target value is the quantity and quality of protein expression within the biological production system. In order to reflect this target value within the objective function, we measured two optical signals from each sample, namely the optical density at 280 nm, and 620 nm (OD_{280}, and OD_{620}, respectively). These are proportional to the concentrations of the total protein, and holo-protein, respectively. Thus, the ratio (OD_{620}/OD_{280}) is a quality index ranging from 0 (when there is no holo-protein in the sample) to 1 (when the sample contains pure holo-protein). Yet, in order to amplify the high protein quantity target, the ratio was multiplied by OD_{620}. Thus, our objective function, subject to maximization, was defined by the equation

$$F_{exp} = \frac{OD_{620}}{OD_{280}} * OD_{620} = \frac{{OD_{620}}^2}{OD_{280}} \rightarrow \max \qquad (1)$$

The next stage of applying optimization to a biological experimental system was to define the array of control parameters, which constitute the decision variables. In the case of the ApcA production system, the parameters were growth and expression conditions, as presented in Table 1. The selection was according to the experimental protocols that are commonly used in our laboratory. The ranges of values for each parameter included the values used in the common laboratory protocol. Each combination of parameters was experimentally tested twice, thereby leading to duplicate biological samples.

Table 1. ApcA production system's decision variables

VARIABLE NAME	**PARAMETER**	**VALUES**
T_1	Growth temperature (°C)	25,30,35,40
T_2	induction temperature (°C)	20,25,30,35,40
V	Growth volume (ml)	20,100,200
C_1	IPTG concentration (mM)	0.5,0.8,1,1.3,1.5,0
T_1	Induction timing (OD)	0.4,0.6,0.8,1,1.2
T_2	Induction length (hr)	4,20,24,48
C_2	Γ-ALA concentration (mM)	0,0.5,1,1.5,2,2.5,3
T_3	Γ-ALA adding timing (stage)	Growth, Expression
C_3	$FeCl_3$ concentration in the medium (mM)	0,0.5,1,1.5,2,2.5,3,3.5,4,4.5,5
M	Medium type (according to standarts)	LB, TB, DNB*

* LB (Lysogeny Broth): Difco prepared powder, 25 g/l (contain tryptone 10g/l, yeast extract 5 g/l and sodium chloride 5 g/l); TB (Terrific Broth): Difco prepared powder, 47.6 g/l and 4ml glycerol (contain Pancreatic Digest of Casein 12g/l, Yeast Extract 24g/l, Dipotassium Phosphate 9.4g/l and Monopotassium Phosphate 2.2g/l); DNB (Diluted Nutrient Broth): D+ glucose 1 g/l, sodium chloride 6 g/l, peptone 15 g/l and yeast extract 3 g/l.3.2 Cloning genes for heterologous expression of phycobiliproteins in *E. coli*.

The Evolution Strategy, executing the combinatorial optimization iterative process, was implemented in MATLAB by Dr. Ofer Shir. It was manually run per each laboratory cycle.

3.2 Engineering *E. coli* bacteria for ApcA expression

The system for phycobiliprotein production developed in the laboratory of Prof. Kai-hong Zhao, Huazong Agricultural University, Wuhan, China has been established in our laboratory as part of a collaboration. In this study, the system was adapted to quick expression, cleaning and isolation of ApcA. The cells chosen for protein expression were *E. coli* bacteria type BL21(DE3). These are standard competent cells engineered to receive foreign DNA. Two plasmids containing the T7 promotor were used for overexpression of several proteins necessary for producing the target protein in the bacterial cells. The plasmids included different antibiotic resistance genes as means to ensure selection and transformation, and were engineered as follows:

a. Two genes required for phycocyanobilin's (PCB's) production, namely pcyA and ho1, were cloned into the pACYC-Duet plasmid, which has resistance to chloramphenicol antibiotics.

b. The apcA gene taken from the genomic sequence of Anabaena Sp. Strain PCC 7120 cyanobacteria was extended by 21 DNA nucleotides that encode to 6 histidine amino-acids at its N-terminal for easy separation, isolation and cleaning of the protein by nickel ions affinity. This gene was cloned together with the Lyase-CpcS gene responsible for chromophorilation with PCB into the pCDF-Duet plasmid that has resistance to streptomycin, and kanamycin antibiotics.

Bacteria were transformed by heat shock, and were tested by growing on LB + agar medium in petri dishes containing the three antibiotics. Bacterial colonies that grew on the agar medium were isolated into liquid LB medium containing three antibiotics and 30% glycerol, frozen, and stored at -80°C.

3.3 Bacterial growth and ApcA production

Transformed bacteria containing the genes for ApcA expression were grown at different conditions according to the optimization algorithm. All growth media and flasks were sterilized prior to growth by autoclave for 20 minutes in 121^{o}C, and contained 35 mg/l chloramphenicol, 50 mg/l kanamycin and 100 mg/l streptomycin.

An initial starter culture of the bacteria was grown in a tube containing 20 g/l LB medium. The volume of the starter was determined according to the growth volume and dilution factors, which were explicit decision variables. The starter culture was incubated with shaking overnight at 37°C, and then was diluted 100 fold by transferring into a flask containing the growth medium. The flask was incubated with shaking at different growth temperatures (T_1), while checking the optical density (OD) at 600 nm until it reached its prescribed value (t_1).

The induction phase was initiated by adding isopropyl β-D-thiogalactopyranoside (IPTG) to the culture at various concentrations (C_1). Then, the culture was incubated with shaking at different induction temperatures (T_2), and durations (t_2). At the end of the induction stage, the bacteria cultures were centrifuged at 4°C, 8000 rpm for 10 minutes, and were stored at -20°C until further processing. Bacterial cells were lysed by suspending their pellets in a lysis buffer comprising 0.02M phosphate buffer, 0.5M NaCl 1 mg/ml of lysozyme. The suspension was incubated at room temperature (22-25°C) for 30 minutes for effective activation of lysozyme. After incubation, the suspension was subject to sonication (two cycles of 1.5 minutes at 20kHz). For breaking the cells, and to ensure absence of DNA, 10U benzonase were added, and the sample was incubated on ice for 20 minutes. This was followed by high-speed centrifugation (12000 rpm) at 4°C for 20 minutes to separate the cell debris from the lysate containing the soluble ApcA that was kept for further processing.

3.4 High-throughput ApcA separation and chromophorylation assay

For ApcA separation and cleaning, we chose to implement nickel affinity chromatography using the His MultiTrap 96-well plates (GE healthcare). Each well in the plate contains a column of chromatography medium with Ni^{2+} ions that bind specifically and with high-affinity to specific six-histidine tag located at the edge of the protein's N-terminal. This method is chosen for its many advantages, particularly easy implementation protocol, and the option to rapidly and simultaneously clean and separate up to 96 samples. The first stage of cleaning was washing the columns with a washing buffer containing 0.02M Phosphate Buffer and 0.5M NaCl, followed by centrifugation. This stage was done twice to ensure complete wash of the columns. In the next stage, samples of cell lysates were loaded onto the nickel columns, and the columns were washed twice again by centrifugation as in the first stage in order to remove all the unbound proteins. Finally, ApcA was eluted from each column with a washing buffer containing 0.5M imidazole. The eluted samples were collected in UV-Vis transparent 96-well plate and transferred to plate reader (TECAN, infinite M200PRO) for recording the absorption spectra of each sample between 250-750 nm. The raw digital data was processed by a specific scripting package written by Dr. Dror Noy for the 'Igor 7' data analysis software. Data processing included subtracting a reference spectrum of the elution buffer, recorded from two wells in another plate, from each sample's spectrum, extracting the absorption values at 280 nm and 620 nm from each corrected sample spectrum and using these as the OD_{280} and OD_{620} signals for determining the objective function values. Further statistical analyses and plotting were performed using scripts written by Mr. Assaf Israeli in R-Studio.

4. Results

4.1 Evolution of target values and decision variables

The optimization algorithm produced in every iteration ("cycle") 24-candidate solutions for maximizing the ratio between the ApcA holoprotein and apoprotein. Each candidate is a combination of the 10 prescribed external decision variables. The 24x10 matrix of candidate combinations is called a "generation". It is based on the combinations with the highest target values of the previous generation, with the exception of the initial set, which is uniformly generated at random. This way, the algorithm iteratively attempts to propose combinations with higher target values. Initially, every combination was produced in duplicates, one containing IPTG as prescribed by the optimization algorithm, and another without IPTG as a control. Yet, since we noticed significant levels of protein expression in some of the control samples, we decided to include a value of 0 (no IPTG) in the range of IPTG concentrations (adhering to the C_1 variable), and give up the control. Consequently, since the control samples became valid combinations, the first generation contained 48 combinations in practice.

Fig. 5 shows the development of target values throughout all the tested generations - six in total. As expected, the first generation yielded low target values because of the random initialization, but these improved in the following generations. The top target value per generation increased as the algorithm progressed, improving from 0.03 in the first generation to 0.21 in the one before last generation, while dropping slightly to 0.13 in the last generation. In addition to the increasing target values, the proportion of high quality combinations increased from generation to generation, and less combinations resulted in target values close to zero as the optimization process progressed. In the first generation, 17 out of 48 combinations yielded extremely low target values, but in the last generation, only 4 combinations out of 24 yielded target values close to zero. Fig. 6 depicts all 168 combinations tested so far in a gallery of bar graphs. Each graph presents the values of the decision variables colored according to the target values. The top 25% target values overall are presented in Fig. 7A, and the corresponding subsection of the gallery showing the combinations yielding these top 25% target values are shown in Fig. 7B. Evidently, the top four target values are from the last two generations (colored green and yellow), and are two- to five-fold higher than the rest of the top 25% values. Most of the top combinations present high similarity in the decision space, with small variations (e.g., single modifications). The highest target value was obtained in the 5th generation and constitutes yield of 21%. Table 2 reveals significant differences between the decision variables of the best-attained combination by the algorithm and the best-practiced protocol that was used heretofore.

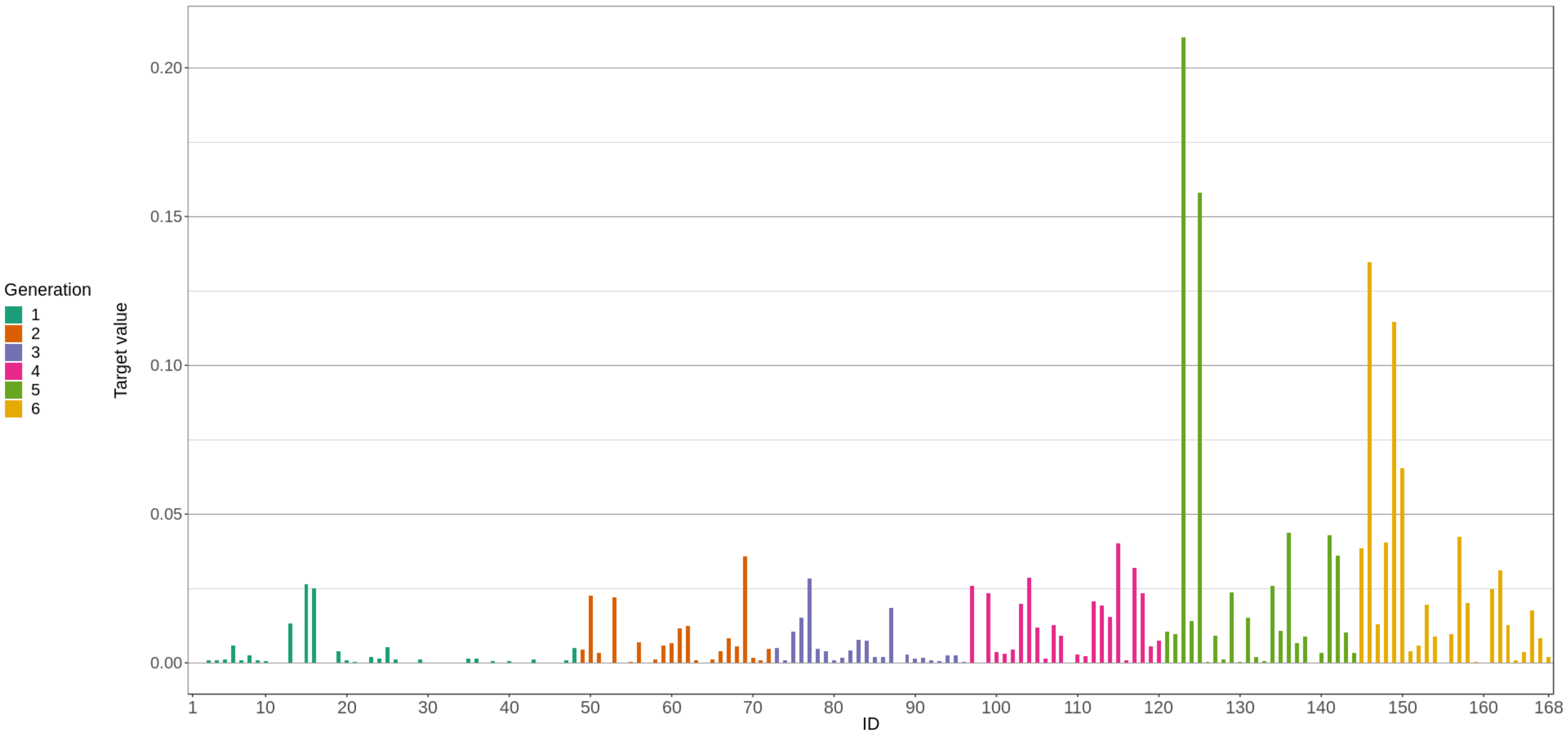


Figure 5. Evolution of target values**.** Bar graph of target values grouped and colored by generation.

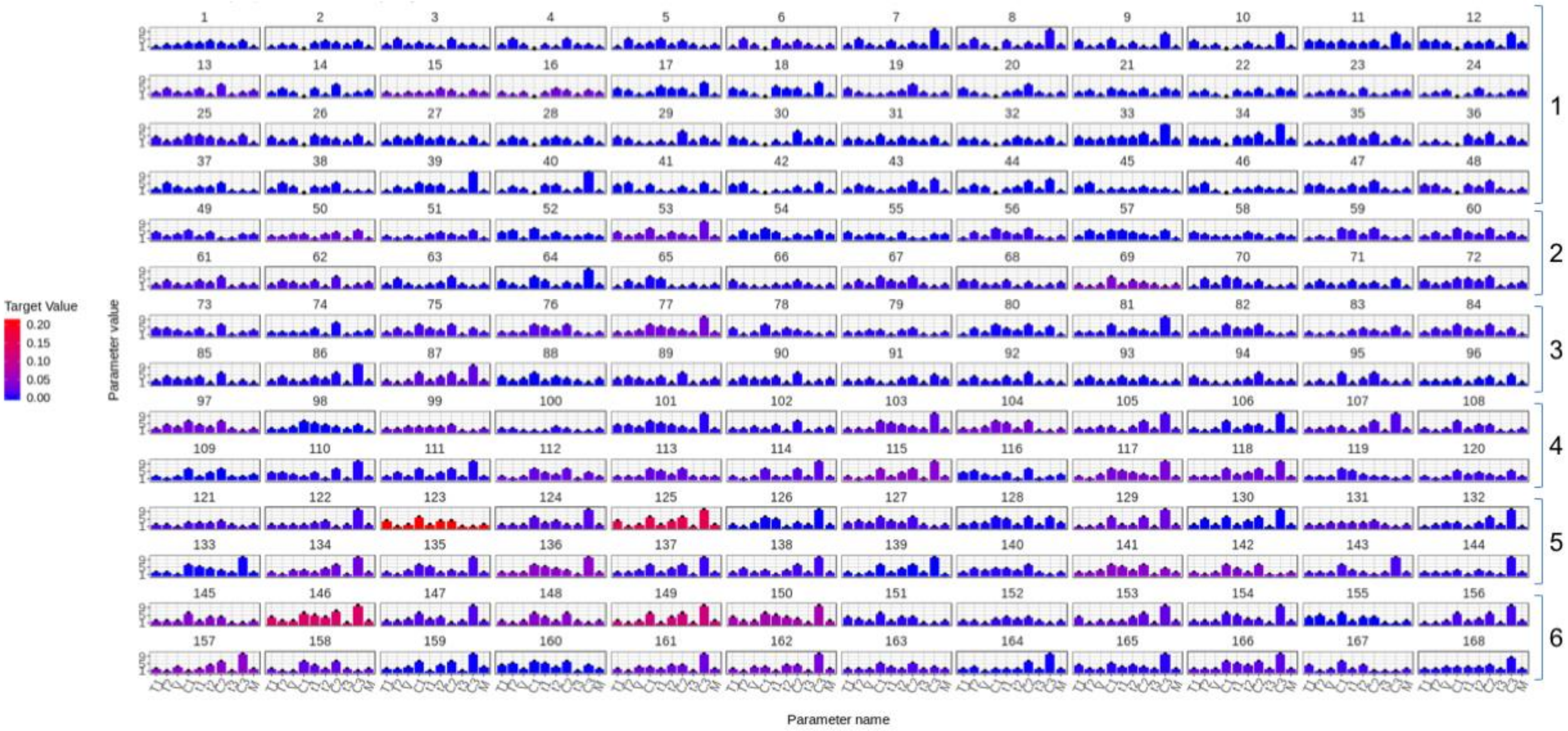


Figure 6. A gallery of all the evaluated combinations throughout the experimental campaign, depicting the explicit decision variables' values in bar-plots. Each combination is identified by its serial number (top of the bar-plot). Within each bar-plot, the X-axis describes the decisions variables and the Y-axis shows their values. The color-map reflects the objective function values adhering to the color legend on the left side of the chart. Generation numbers are indicated on the right side.

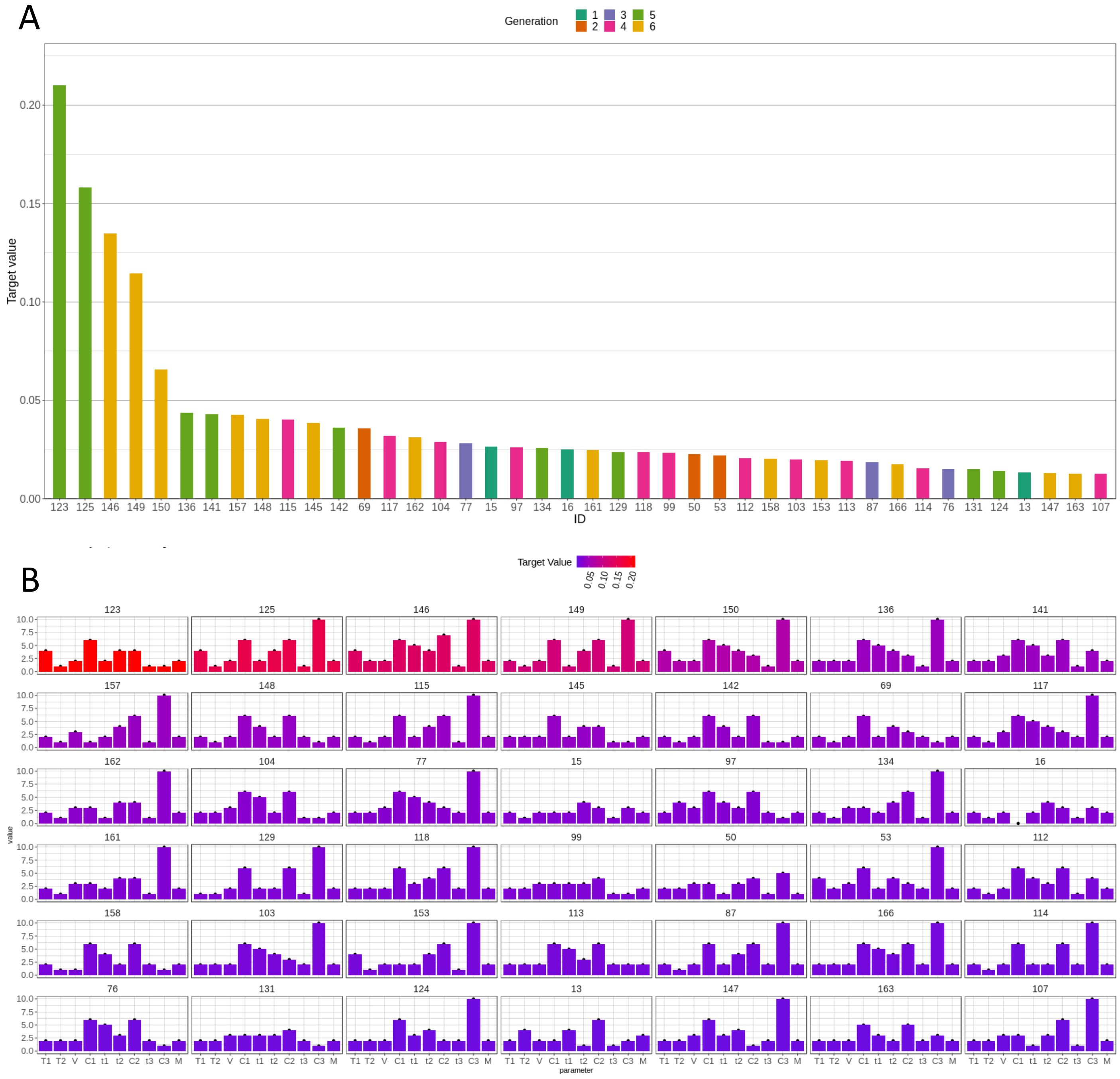


Figure 7. Top 25% combinations in terms of target values. **A.** Bar graph of the top 25% target values ordered from high to low values, colored by generation. **B.** A gallery of the combinations of the top 25% objective function values throughout the experimental campaign, depicting the explicit decision variables' values in bar-plots. Each combination is identified by its serial number with respect to the entire process (top of the bar-plot). Within each bar-plot, the X- axis describes the decision variables and Y-axis shows their values. The color-map reflects the objective function values adhering to the color legend in the right hand side of the chart.

Table 2. Best-attained combination by the algorithm vs. best-practiced protocol

Variable	Parameter	Best-Practiced Protocol	Best-Attained by Algorithm
T_1	Growth temperature (°C)	37^{o}C	30^{o}C
T_2	induction temperature (°C)	20^{O}C	20^{O}C
V	Growth volume (ml)	1000	200
C_1	IPTG concentration (mM)	1	0
t_1	Induction timing (O.D)	0.6-0.8	1.2
t_2	Induction length (hr)	12-18	48
C_2	Γ-ALA concentration (mM)	0	3
t_3	Γ-ALA adding timing (stage)	-	growth
C_3	$FeCl_3$ concentration in the medium (mM)	0	4.5
M	Medium type (according to standards)	LB	TB

4.2 Dependencies among decision variables

In this study, the chosen decision variables, i.e. the growth and expression conditions described in table 1, were external parameters. To examine if there are some dependencies between pairs of parameters, we conducted a basic correlation analysis. This statistical method assigns a correlation coefficient to each pair of variables in an experimental series that reflects the type and the degree of relationships between them. Correlation coefficients approaching +1 imply strong pairwise "alignment", meaning that increasing one variable's values necessarily occurs with increasing the other's. Correlation coefficients approaching -1 imply strong pairwise "anti-alignment" meaning that increasing one variable's values necessarily occurs with decreasing the other's. Correlation coefficients approaching 0 indicate poor alignment in either direction. Fig. 8 presents the pairwise correlation analysis between the decision variables in all tested generations. All the correlation coefficients are below 0.5 with the exception of the correlation between t2 (time of expression) and T2 (temperature of expression) that equals 0.5. These values indicate weak correlation among the decision variables, which means that each decision variables have different effect on the target values independent of any other variable. Importantly, this analysis was carried out using the generated set of candidate solution, being a biased distribution of the search process, which clearly does not represent the general distribution of points in the search space.

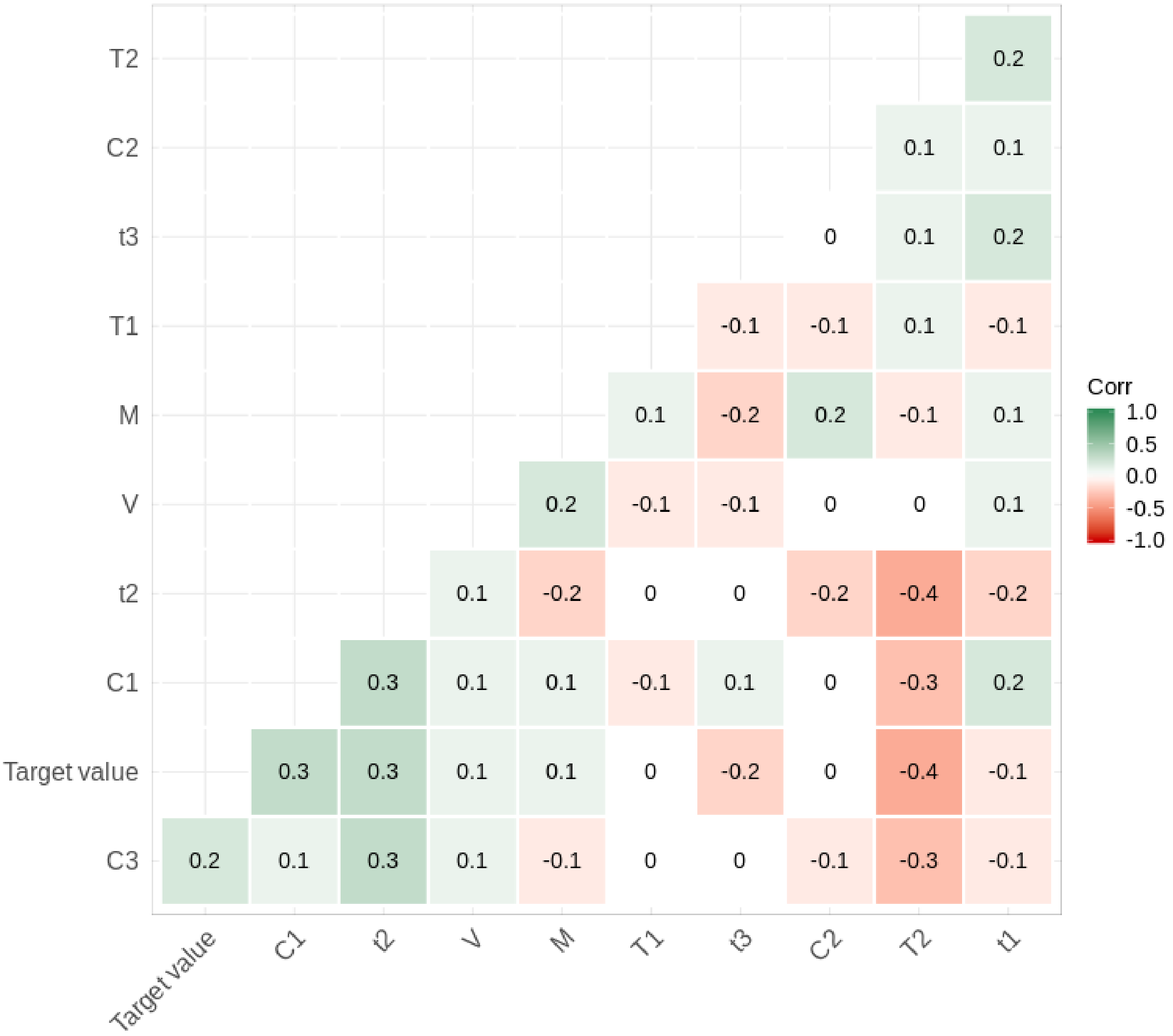


Figure 8. Pairwise correlations between decision variables. Correlation coefficient values presented in circles at the crossing of decision variable pairs from the X- and Y- axes. Variables labeled by their short names (materials and methods, 3.1). The circles size and color scaled according to each coefficient value and sign.

5. Discussion

In this work, we implemented a combinatorial optimization methodology to address the challenge of achieving maximal levels of the phycobiliprotein ApcA in complex with its native chromophore, PCB, by heterologous expression in *E. coli*. We observed a change in the tendency, which includes improvement of the expression and production of the holo-protein, but the results are yet lower than what we had expected. Further analysis of the results can lead us to a new direction in future research. In the following, we discuss the effect of the objective function formulation, as well as the insights concerning the biological aspects.

5.1 Alternative objective function definition

The objective function was intuitively defined to achieve two purposes: 1) increasing the overall quantity of the protein, and 2) maximizing the yield of chromophorylated protein. As explained in detail in 4.1, these considerations were mathematically formulated as the ratio between the squared values of OD_{620}, to OD_{280}. Yet, after five cycles, the chromophorylated protein yields remained below 21%. A possible reason for the slow improvement in production yield is that the current objective function definition represents the experimental goals only to a limited degree. To better understand the properties of the current objective function, we compared the evolution of each of its components, namely, OD_{620}, and OD_{620}/OD_{280} throughout the optimization campaign. Figure 9 presents the objective function values, OD_{620}, and OD_{620}/OD_{280} for the second, fifth and sixth generations. Notably, while the OD_{620}/OD_{280} ratios slightly increase and then decreases, the OD_{620} values increase dramatically. This implies that the algorithm compensates for decreasing the OD_{620}/OD_{280} ratio by increasing the total ApcA production. Apparently, this form of the objective function allows maximizing the total quantity of holo-ApcA by increasing total protein production, but without necessarily improving the quality of the preparation. This underlines the importance of defining an adequate objective function for the optimization problem at hand, since it drives the algorithm's search path. In our case, a more adequate alternative to the objective function may be a split (conditional) form, whereby the total ApcA levels are considered up to a *threshold* OD_{620} value, while $O.D_{620}/O.D_{280}$ ratios are considered above the *threshold*, denoted as θ:

$$F_{exp2} = \begin{cases} \frac{{OD_{620}}^2}{OD_{280}} & \textbf{if } \frac{OD_{620}}{OD_{280}} < \theta \\ \frac{OD_{620}}{OD_{280}} & \textbf{otherwise} \end{cases} \rightarrow \max$$

The advantage of this alternative objective function is that it divides the optimization procedure into two main stages in a single run - first, calibration and improvement of total production yield, and second, increasing the yield of the desired molecule (holo-protein ApcA).

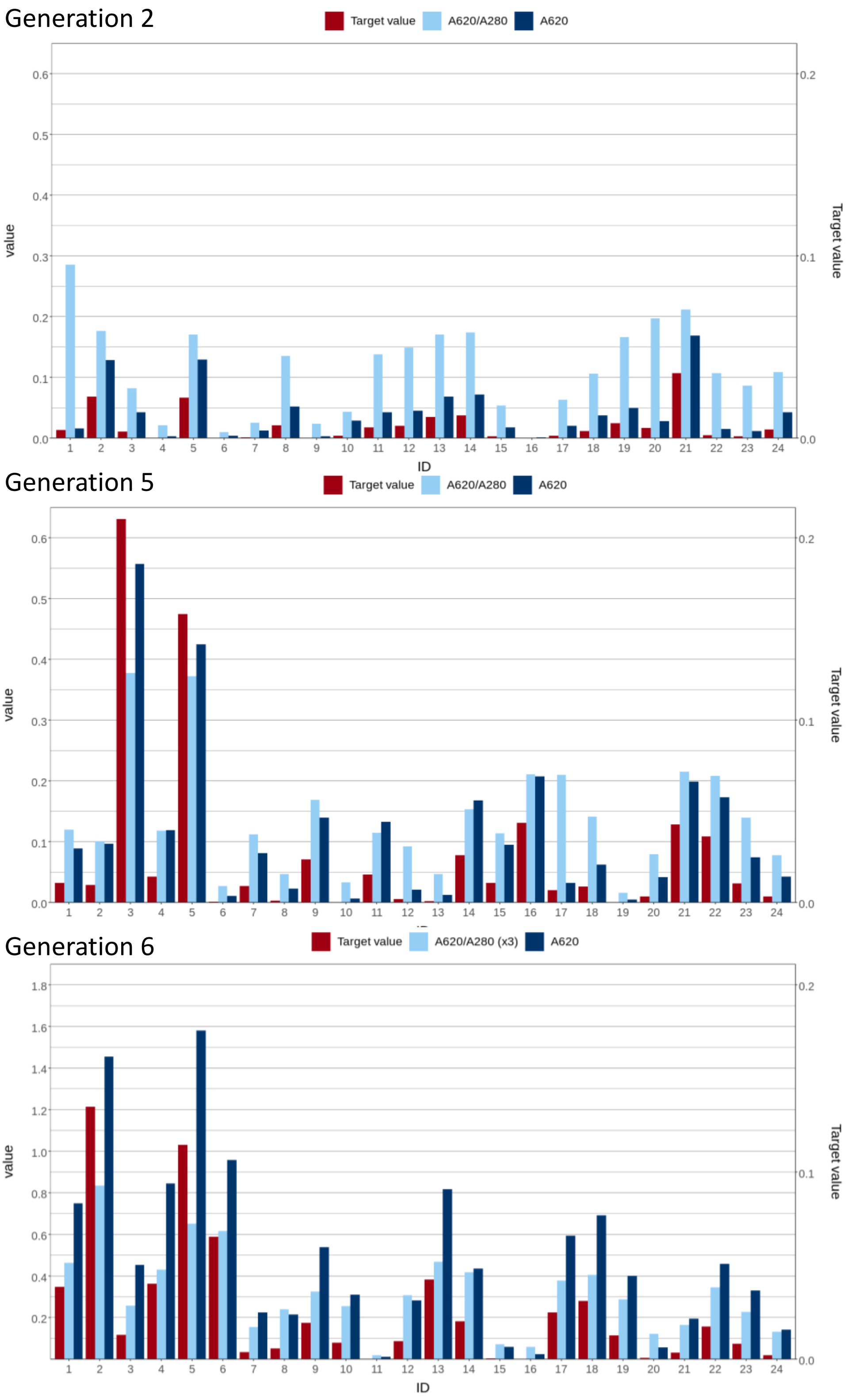


Figure 9. Target values and the components of the objective function for the second, fifth, and sixth generations**.** Bars representing the OD_{620}/OD_{280} ratios, the absolute OD_{620}, and the objective function values are colored cyan, blue, and red, respectively, are shown for each parameter combination sorted by the sample IDs. Y-axis at the right hand side describes the target value while Y-axis at the left hand side shows the values of both components. In generation 6, the scale is three times bigger than in previous generations, so the lower values (ratio $O.D_{620}/O.D_{280}$) were multiplied by a factor of three.

5.2 Insights into the biological and biochemical determinants of phycobiliprotein expression

An important benefit of using a computationally guided search of a broad range of experimental conditions is the possibility of gaining non-intuitive new insights into the biology and biochemistry of the system of interest. To this end, we compared the best-practiced protocol of phycobiliprotein expression to the best-attained protocol according to the computational algorithm (Table 3). Evidently, there are significant differences between the protocols in several parameters including growth temperature, type of growth media, and addition of IPTG, $FeCl_3$, and γ-ALA. The growth temperature varies from 37^{o}C in the best-practiced protocol to 30^{o}C in the best-attained protocol. While the lower temperature is still within the growth range of *E. coli*, a mesophilic bacterium capable of growing between 28^{o}C from 40^{o}C[30], the usual growth temperature for protein production applications is 37^{o}C. This is common knowledge based on the fact that *E. coli* is a gut bacterium hence the temperature of its natural environment is 37^{o}C, the mean body temperature of most mammals[31]. Finding an optimal growth temperature of 30^{o}C for our specific protein production system is a nontrivial task. It demonstrates the advantage of a computational algorithm that uses an unbiased, conception-free, systematic search, over a more intuitive, conception-driven and empirical search of protein expression conditions.

Table 3 shows the differences in the components of the LB, and TB growth media used in the best-practiced, and best-attained protocols, respectively. Clearly, TB is a richer medium and contains higher percentage of nutrients providing energy sources for different metabolic processes. In addition to the richer growth medium, the best-attained protocol includes high concentrations of $FeCl_3$ and γ-ALA, but excludes IPTG. The latter is an effective inducer of protein expression in *E. coli* when the genes to be expresses are under the control of the lac operon. This is one of the most commonly used systems for protein production using *E. coli*[32]. Its omission in the best-attained protocol is counter-intuitive. In hindsight, the explanation for this result is that we used the T7 RNA polymerase promoter. This is one of the most common promoters for protein expression in *E. coli*, but it is also known for its “leaky” expression – significant levels of expression in the absence of inducer, when used in multiple plasmid systems[30]. In our case, reducing the expression levels of the apoproteins is beneficial because pigment biosynthesis and chromophorylation should keep up with protein synthesis. The expression levels in the absence of IPTG are much lower than in its presence, but are still enough to provide significant protein expression.

Table 3. Commonly used components of types of culture media[33]

Medium	Amount of component							
	Yeast extract (%)	Tryptone (%)	Soytone (%)	NaCl (%)	KH_2PO_4/K_2HPO_4 (mM)	Glycerol (%)	Total amino nitrogen (%)	Total carbohydrate (%)
LB	0.5	1	-	0.5	-	-	0.9	0.1
TB	2.4	1.2	-	-	90	0.4	3.4	0.4

In contrast to IPTG concentrations, the decision to test adding $FeCl_3$ and γ-ALA to the growth media, was based on the previous knowledge that heme are precursors of phycocyanobilins, and that iron and γ-ALA are essential for heme biosynthesis. Particularly, γ-ALA is the rate-limiting factor of the initial step in the pathway of heme biosynthesis.[34] Thus, addition of exogenous iron and γ-ALA can boost heme synthesis in *E. coli* and should increase the production of PCB[35]. This prediction was validated by the fact that the best-attained protocol includes addition of $FeCl_3$ and γ-ALA to the growth medium.

6. Conclusions

This study tested the feasibility of utilizing computational optimization algorithms for improving the heterologous expression yield of holo-ApcA in an *E. coli* system. Although the applied algorithm accomplished some improvement in production yield, it was still far from maximizing the ratio between the holoprotein to the apoprotein. An obvious reason for this is that only six cycles were carried out in the timeframe of this project, which is most likely insufficient for the search to converge. Nonetheless, we were able to gain valuable insights into how to set up an experimental optimization campaign. A key point is the proper definition of the objective function. In this campaign, we used an intuitive definition for the objective function, which proved to bias the computational search toward maximizing total protein production instead of improving protein quality. Defining an adequate objective function that 'translates' intuition into mathematical formulae is critical for a successful experimental optimization campaign. This is a major challenge since, by definition, the targeted system cannot be properly simulated by computational models. Developing computational tools for assessing objective functions in experimental optimization, even partially, is therefore important for developing practical experimental optimization routines.

A useful aspect of the algorithm-based learning system used in this study is the insights into the biological system provided by comparing the best-attained protocol to the best-practiced one, which may lead to new discoveries, even counter-intuitive, with a potential to improve protein production. One of these discoveries is that IPTG, an effective inducer, is not necessary when using the T7 promotor. This insight gives us a hint for future similar experiments to using more regulated expression system with high synchronization between genes to get more accurate target values.

We cannot infer that the algorithm obtains the best combination of parameters for protein production, probably because we applied a small number of iterations in the manner of time we had, but we did learn how to calibrate the experimental combinatorial method for our specific heterologous expression system. Future work in this domain should remain interdisciplinary in nature - conduct longer experimental campaigns on one hand, and attempt at investigating the combinatorial search space on the other.

7. References

1. De Wildt, R. M. T., Mundy, C. R., Gorick, B. D. & Tomlinson, I. M. Antibody arrays for high-throughput screening of antibody-antigen interactions. *Nat. Biotechnol.* **18,** 989–994 (2000).
2. White, R. E. High-Throughput Screening in Drug Metabolism and Pharmacokinetic Support of Drug Discovery. *Annu. Rev. Pharmacol. Toxicol.* **40,** 133–157 (2000).
3. Papadimitriou, C. H. & Steiglitz, K. *Combinatorial Optimization: Algorithms and Complexity*. (Dover Publications, 1998).
4. Cook, W. J., Cunningham, W. H., Pulleyblank, W. R. & Schrijver, A. *Combinatorial Optimization*. (John Wiley and Sons, 2011).
5. Shir, O. M. & Bäck, T. Sequential Experimentation by Evolutionary Algorithms. in *Proceedings of the Genetic and Evolutionary Computation Conference Companion* 956–976 (ACM, 2018). doi:10.1145/3205651.3207885
6. Calzolari, D. *et al.* Search algorithms as a framework for the optimization of drug combinations. *PLoS Comput. Biol.* **4,** (2008).
7. Grossman, A. R., Schaefer, M. R., Chiang, G. G. & Collier, J. L. The Phycobilisome , a Light-Harvesting Complex Responsive to Environmental Conditionst. *Microbiol. Mol. Biol. Rev.* **57,** 725–749 (1993).
8. Zhao, K.-H. *et al.* Phycobilin:cystein-84 biliprotein lyase, a near-universal lyase for cysteine-84-binding sites in cyanobacterial phycobiliproteins. *Proc. Natl. Acad. Sci.* **104,** 14300–14305 (2007).
9. Scheer, H. & Zhao, K. H. Biliprotein maturation: The chromophore attachment. *Molecular Microbiology* **68,** 263–276 (2008).
10. Oi, V. T., Glazer, A. N. & Stryer, L. Fluorescent phycobiliprotein conjugates for analyses of cells and molecules. *J. Cell Biol.* **93,** 981–986 (1982).
11. Santiago-Santos, M. C., Ponce-Noyola, T., Olvera-Ramírez, R., Ortega-López, J. & Cañizares-Villanueva, R. O. Extraction and purification of phycocyanin from Calothrix sp. *Process Biochem.* **39,** 2047–2052 (2004).
12. Jespersen, L., Strømdahl, L. D., Olsen, K. & Skibsted, L. H. Heat and light stability of three natural blue colorants for use in confectionery and beverages. *Eur. Food Res. Technol.* **220,** 261–266 (2005).
13. Sekar, S. & Chandramohan, M. Phycobiliproteins as a commodity: Trends in applied research, patents and commercialization. *Journal of Applied Phycology* **20,** 113–136 (2008).
14. Rimbau, V., Camins, A., Romay, C., González, R. & Pallàs, M. Protective effects of C-

phycocyanin against kainic acid-induced neuronal damage in rat hippocampus. *Neurosci. Lett.* **276,** 75–78 (1999).

15. Liu, Y., Xu, L., Cheng, N., Lin, L. & Zhang, C. Inhibitory effect of phycocyanin from Spirulina platensis on the growth of human leukemia K562 cells. *J. Appl. Phycol.* **12,** 125–130 (2000).
16. Romay, C. *et al.* Antioxidant and anti-inflammatory properties of C-phycocyanin from blue-green algae. *Inflamm. Res.* **47,** 36–41 (1998).
17. Romay, C., Gonzalez, R., Ledon, N., Remirez, D. & Rimbau, V. C-Phycocyanin: A Biliprotein with Antioxidant, Anti-Inflammatory and Neuroprotective Effects. *Curr. Protein Pept. Sci.* **4,** 207–216 (2003).
18. Nagaoka, S. *et al.* A Novel Protein C-Phycocyanin Plays a Crucial Role in the Hypocholesterolemic Action of Spirulina platensis Concentrate in Rats. *J. Nutr.* **135,** 2425–2430 (2005).
19. Cherng, S. C., Cheng, S. N., Tarn, A. & Chou, T. C. Anti-inflammatory activity of c-phycocyanin in lipopolysaccharide-stimulated RAW 264.7 macrophages. *Life Sci.* **81,** 1431–1435 (2007).
20. Bhat, V. B. & Madyastha, K. M. C-Phycocyanin: A potent peroxyl radical scavenger in vivo and in vitro. *Biochem. Biophys. Res. Commun.* **275,** 20–25 (2000).
21. Soni, B., Trivedi, U. & Madamwar, D. A novel method of single step hydrophobic interaction chromatography for the purification of phycocyanin from Phormidium fragile and its characterization for antioxidant property. *Bioresour. Technol.* **99,** 188–194 (2008).
22. Bermejo, P., Piñero, E. & Villar, Á. M. Iron-chelating ability and antioxidant properties of phycocyanin isolated from a protean extract of Spirulina platensis. *Food Chem.* **110,** 436–445 (2008).
23. Benedetti, S. *et al.* Antioxidant properties of a novel phycocyanin extract from the blue-green alga Aphanizomenon flos-aquae. *Life Sci.* **75,** 2353–2362 (2004).
24. Tooley, A. J., Cai, Y. A. & Glazer, A. N. Biosynthesis of a fluorescent cyanobacterial C-phycocyanin holo- subunit in a heterologous host. *Proc. Natl. Acad. Sci.* **98,** 10560–10565 (2001).
25. Xu, Q. Z. *et al.* Far-red light photoacclimation: Chromophorylation of FR induced α- and β-subunits of allophycocyanin from Chroococcidiopsis thermalis sp. PCC7203. *Biochim. Biophys. Acta - Bioenerg.* **1857,** 1607–1616 (2016).
26. Tang, K. *et al.* The terminal phycobilisome emitter, L_{CM}: A light-harvesting pigment with a phytochrome chromophore. *Proc. Natl. Acad. Sci.* **112,** 15880–15885 (2015).

27. Zeng, X. L. *et al.* Bimodal intramolecular excitation energy transfer in a multichromophore photosynthetic model system: Hybrid fusion proteins comprising natural phycobilin- and artificial chlorophyll-binding domains. *J. Am. Chem. Soc.* **135,** 13479–13487 (2013).

28. Zhang, J. *et al.* Fused-gene approach to photoswitchable and fluorescent biliproteins. *Angew. Chemie - Int. Ed.* **49,** 5456–5458 (2010).

29. Emmerich, M., Shir, O. M. & Wang, H. in (eds. Martí, R., Panos, P. & Resende, M. G. C.) 1–31 (Springer International Publishing, 2018). doi:10.1007/978-3-319-07153-4_13-1

30. Rosano, G. L. & Ceccarelli, E. A. Recombinant protein expression in *Escherichia coli*: Advances and challenges. *Frontiers in Microbiology* (2014). doi:10.3389/fmicb.2014.00172

31. Nguyen, M. T. *et al.* The effect of temperature on the growth of the bacteria *Escherichia coli* DH5 α. *Saint Martin's Univ. Biol. J.* (1936).

32. IPTG Induction Theory, Biologics International Corp.

33. Danquah, M. K. & Forde, G. M. Growth Medium Selection and Its Economic Impact on Plasmid DNA Production. *J. Biosci. Bioeng.* (2007). doi:10.1263/jbb.104.490

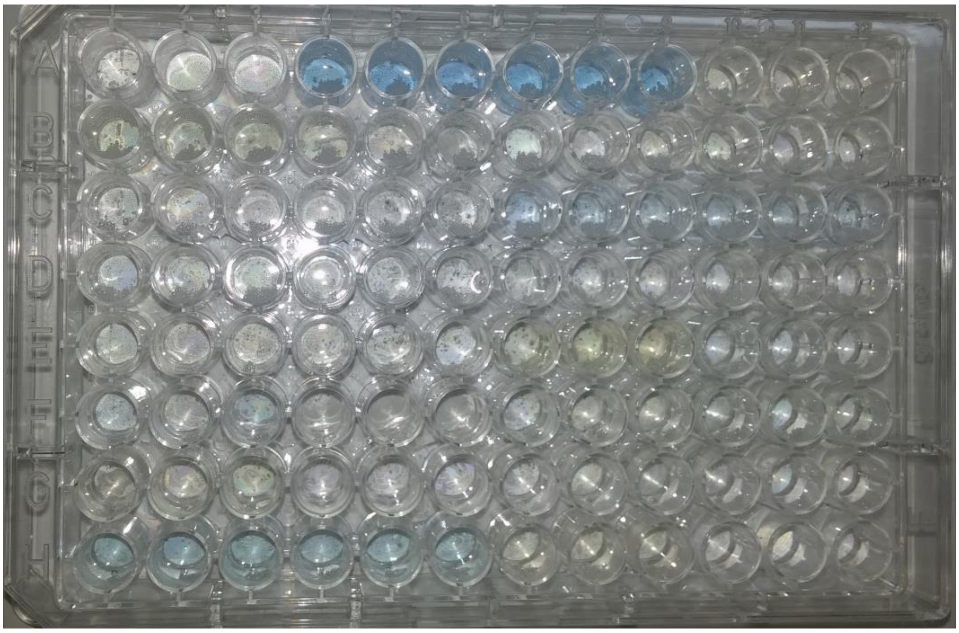
34. Schobert, M. & Jahn, D. Regulation of heme biosynthesis in non-phototrophic bacteria. *J.Mol.Microbiol.Biotechnol.* (2002).

35. Ge, B. *et al.* Combinational biosynthesis of phycocyanobilin using genetically-engineered *Escherichia coli*. *Biotechnol. Lett.* (2013). doi:10.1007/s10529-012-1132-z

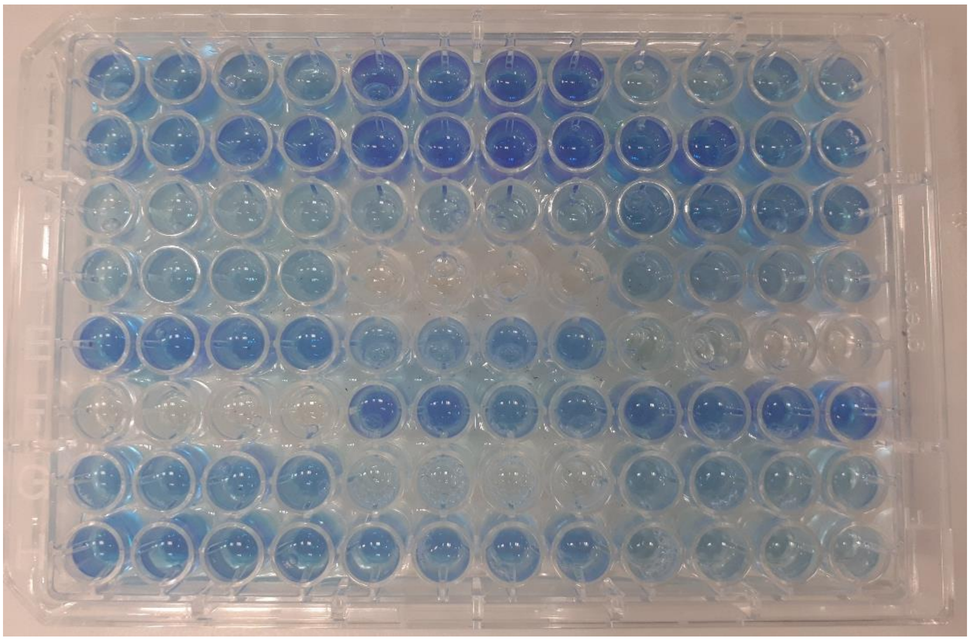
Supplementary Information

Separated proteins collection

Pictures of the first and the 6th generation's collection plate, after cleaning and separation of the proteins. The blue pigment is clear to the human eye and can hint about the target value- as bluer as higher.



Generation 1



Generation 6

הפקולטה למדעים

תכנית מוסמך מחקרי בביוטכנולוגיה

נושא העבודה בעברית : אופטימיזציה קומבינטורית ניסיונית של ביטוי פיקוביליפרוטאינים בחיידקי *E. coli*

נושא העבודה באנגלית:

Experimental combinatorial optimization of phycobiliproteins' expression in *E. coli*

עבודת גמר לשם מילוי חלקי של הדרישות לקבלת התואר "מוסמך מחקרי בביוטכנולוגיה"

מאת

חן ארליך

Chen Erlich

בהנחיית:

ד"ר דרור נוי וד"ר עפר שיר, מיג"ל- מרכז ידע גליל עליון

תאריך : פברואר 2019

תקציר

מבחני *High throughput screening (HTS)* משמשים במערכות ביולוגיות בעלות מורכבות גבוהה הזקוקות לחיפוש במרחב הכולל טווח רחב של פרמטרים. במקרים כאלה, גישות אופטימיזציה אשר מבצעות חיפוש על פרמטר יחיד או תת-קבוצה של פרמטרים, אינן יעילות. לכן, הגדרת מדידה בעלת רמת דיוק גבוהה, מהירות, יעילות, רגישות וצריכה נמוכה של מגיבים – הינה קריטית ליישום מוצלח של HTS. גישה חדשנית לתכנון מבחני HTS רואה במרחב החיפוש כבעיית אופטימיזציה קומבינטורית ומבצעת חיפוש מונחה-אלגוריתם הפועל על כל מערך הפרמטרים. ברוב המערכות הביולוגיות, היות ואין אפשרות לקבוע באופן אנליטי או לדמות את התלות של ערך פונקציית המטרה בפרמטרים הניסיוניים, נדרשת מדידה לצורך קביעת ערך המטרה בפועל. סוג זה של בעיות אופטימיזציה משתייך לתחום הנקרא אופטימיזציה ניסיונית (experimental optimization). בעבודה זו, אנו מיישמים אופטימיזציה ניסיונית במערכת ביטוי הטרולוגית של Allophycocyanin A (ApcA), פיקוביליפרוטאין ציאנובקטריאלי בחיידקי *E. coli*. פיקוביליפרוטאינים הם חלבונים מסיסים במים אשר קושרים קוולנטית לשייר ציסטאין ספציפי מולקולות פיגמנט טטרה-פירוליות לינאריות (כרומופורים) הנקראות פיקובילינים (phycobilins). בציאנובקטריה ואצות אדומות, החלבונים קושרי הכרומופורים יוצרים יחד עם חלבונים נוספים נטולי כרומופורים קומפלקסים עצומים הידועים כפיקוביליזומים. הארגון המרחבי של תת-היחידות של הפיקוביליפרוטאינים בפיקוביליזום כרשת משוכללת של קומפלקסים של חלבון-פיגמנט מספק לאורגניזמים הפוטוסינתטיים מערכת חזקה ודינמית לקליטת אור השמש והעברת האנרגיה לשם ניצולה בתהליכים מטבוליים. מלבד תפקידם כאבני הבניין של קולטני אור פוטוסינתטיים, יש עניין רב בפיקוביליפרוטאינים כסמנים ספקטרוסקופיים ביישומים רפואיים וביוטכנולוגיים.

ApcA קושר כרומופור יחיד (PCB=phycocyanobilin) והוא חלק מתת-היחידות המרכיבות את ליבת הפיקוביליזום. מערכת ביטוי היתר ההטרולוגית שלו כוללת ארבעה גנים: אחד לביטוי ספציפי של החלבון, שני גנים אשר מעורבים בביוסינתזה של PCB, והגן הרביעי מבטא אנזים מסוג lyase אשר מזרז את הקישור הקוולנטי הספציפי של הכרומופור לחלבון. מערכת כזו מתאימה מאוד לאופטימיזציה ניסיונית מכיוון ש: 1) יעילותה תלויה במספר רב של פרמטרים, חלקם פרמטרים חיצוניים הנשלטים בקלות בעוד אחרים הם פרמטרים פנימיים הקשים לשליטה; 2) ספקטרום הבליעה והפליטה של תוצר המטרה יכול להימדד עם רקע נמוך ויחס גבוה של אות לרעש; 3) התפוקה של קומפלקס החלבון-פיגמנט ניתנת להגדרה כפונקציית מטרה; 4) הפרוטוקולים הניסיוניים מבוססים, ובאופן יחסי אינם יקרים או צורכים זמן עבודה רב.

במחקר זה, יישמנו אופטימיזציה קומבינטורית ניסיונית במטרה להשיג העלאה בתפוקה של ייצור קומפלקס-ApcA phycocyanobilin רקומביננטי. הגדרנו פונקציית מטרה המבוססת על מדידות אופטיות של דוגמאות החלבון, וכן מרחב חיפוש של עשרה פרמטרים חיצוניים (משתני החלטה) בעלי השפעה על ייצור הקומפלקס. ההגדרות והערכים קודדו באלגוריתם שמאתחל אוכלוסייה אקראית של צירופי פרמטרים ניסיוניים ומפעיל באופן איטרטיבי אופרטורים של וריאציה לבניית אוכלוסיית צאצאים, שמתוכה נבחרת אוכלוסיית ההורים בהתבסס על ערך פונקציית המטרה המתקבל לכל קומבינציה מועמדת. בנינו פרוטוקולים ניסיוניים המקבלים קומבינציה של פרמטרים מן האלגוריתם, יישמנו אותם במערכת ביטוי והפקת חלבון ומדדנו את תוצאות פונקציית המטרה לכל קומבינציה. לאחר מכן, התוצאות הועברו כמשוב-חוזר לאלגוריתם האופטימיזציה ליצירת דורות חדשים של צירופים נבחנים. בסך הכל, נבחנו דור התחלתי שהשווה 48 צירופים וחמישה דורות נוספים שהשוו 24 צירופים כל אחד. גילינו שיפור בערכי פונקציית המטרה מדור אחד לאחר אבל התפוקה המקסימלית

הייתה נמוכה מהמצופה. ניתוח מפורט של התוצאות מגלה שגישת האופטימיזציה מעדיפה הגדלת רמת ביטוי חלבון כללי על פני שיפור היחס בין הקומפלקס חלבון-פיגמנט לסך כל החלבון שיוצר. הדבר מעיד על כך שפונקציית המטרה, אשר הוגדרה בעיקרה על-סמך אינטואיציה, אינה מדייקת בתיאור המטרה. ביישומים עתידיים יש צורך בפיתוח כלים חישוביים להגדרה מתמטית מדויקת יותר של פונקציית המטרה תוך שמירה על הצורך לשקף ביטוי חלבון ספציפי. על אף זאת, המחקר הזה מספק תובנות חדשות, לא-אינטואיטיביות על המערכת הביולוגית. לדוגמא, גילינו כי תחת בקרת פרומוטור T7, התפוקות הגבוהות ביותר מתקבלות ללא הוספת IPTG. תובנה זו מרמזת על שינויים ברמת הביולוגיה המולקולרית שאנו יכולים להכניס בעתיד, המתקבלים מתהליך האופטימיזציה (למשל, החלפת הפרומוטור מ-T7 לפרומוטור בעל יכולת שליטה ובקרה גבוהה יותר).

למרות שהאלגוריתם לא הביא למקסימום את ייצור החלבון בדרגה המצופה, השגנו תובנות הן בפן המתמטי והן בפן הביולוגי של המערכת, מה שמסייע לנו להגדיר ולתכנן בצורה מדויקת יותר את פונקציית המטרה ומשתני ההחלטה על מנת להשיג את היעדים שלנו במחקרים עתידיים.

מילות מפתח: אופטימיזציה קומבינטורית, הפקת חלבונים הטרולוגית, ApcA, Holo-protein