

Metabolic and genetic engineering

Giacomo Fantoni

telegram: @GiacomoFantoni

Github: https://github.com/giacThePhantom/thesis_notes

April 14, 2022

Contents

1	Introduction	2
2	Flux	3
2.1	Introduction	3
2.1.1	Feasibility and observability	3
2.2	Metabolic flux analysis	3
2.2.1	Flux analysis at metabolic steady-state: MFA	3
2.2.2	Flux balance analysis	4
2.2.3	¹³ C-based metabolic flux analysis	4
3	Noise	5
3.1	Regulation of noise in the expression of a single gene	5
3.2	Adaptive response of gene network to environmental changes by fitness-induced at- tractor selection	6
4	CFPS and MAGE	10
4.1	Cell-free protein synthesis	10
4.1.1	Cell extract	11
4.1.2	PURE system	11
4.2	Multiplex Automated Genome Engineering	12
4.2.1	MAGE automation	15
4.3	CFPS and MAGE	15

Chapter 1

Introduction

Chapter 2

Flux

2.1 Introduction

The metabolic flux can be defined as the rate at which material is processed through a metabolic pathway. Along with intracellular metabolite concentrations, fluxes define the minimal information needed for describing metabolism and cell physiology. Metabolic fluxes and their changes in response to various types of genetic and environmental perturbations are critical for the elucidation of the control of metabolic flux.

2.1.1 Feasibility and observability

A metabolic pathway is defined to be any sequence of feasible and observable biochemical reaction steps connecting a specified set of input and output metabolites. If the fluxes of reaction sequences cannot be determined independently, their inclusion provides no additional information. In many ways, it is preferable to lump these reaction sequences together in fewer pathways whose overall fluxes can be experimentally observed.

2.2 Metabolic flux analysis

Metabolic flux analysis or MFA has been used over the past three decades to quantify intracellular metabolic fluxes in native and engineered biological systems. Through MFA changes in metabolic pathway fluxes that result from genetic and or environmental interventions are quantified. This information provides insights into the regulation of metabolic pathways and may suggest new targets for further metabolic engineering of the strains.

2.2.1 Flux analysis at metabolic steady-state: MFA

The first step in an MFA is to express the biochemical network model as a stoichiometric matrix in which rows represent balanced intracellular metabolites and columns represent metabolic fluxes in the model. The stoichiometric model includes a biomass reaction that describes the drain of precursor metabolites needed for cell growth, which is constructed based on the measured biomass composition. By assuming metabolic pseudo steady-state for intracellular metabolites, metabolic fluxes \vec{v} are constrained by the stoichiometry matrix S :

$$S \times \vec{v} = \vec{0}$$

To estimate metabolic fluxes, the stoichiometric constraints are complemented with measured external metabolic rates, such as growth rate, substrate uptake and product accumulation rates, described by matrix R . This adds the constraint:

$$R \times \vec{v} = \vec{r}$$

The combined system of equation is solved by least squares regression:

$$\min SSR = \sum \frac{(r - r_m)^2}{\sigma_r^2}$$

With constraints $R \times \vec{v} = \vec{r}$ and $S \times v = \vec{0}$. MFA can estimate metabolic fluxes in systems fully or over-determined: all the necessary external rate measurement are known or they are redundant. This method is easy to apply and relies on relatively robust measurements of extracellular metabolites. However for many biological systems the number of constraints is often insufficient to observe all important metabolic pathways. To make the system fully observable additional assumptions are needed. For example specific pathways that are assumed to carry little or no flux or cofactor balances are left out.

2.2.2 Flux balance analysis

Flux balance analysis or FBA can be applied to quantify fluxes in underdetermined systems. In addition to applying constraints from measured extracellular rates, inequality constraints like upper and lower bounds on fluxes are used and an assumed biological objective is imposed on the model. FBA returns a large solution space consisting of many flux distributions that can all maximize the assumed cellular objective.

2.2.3 ^{13}C -based metabolic flux analysis

^{13}C -based metabolic flux analysis or ^{13}C -MFA is a more advanced technique for estimating metabolic fluxes in systems at metabolic steady state. ^{13}C -labelled tracers, combined with isotopomer balancing, metabolite balancing and isotopic labelling measurement through techniques as NMR, mass spectrometry and tandem mass spectrometry are used to estimate fluxes. Cells are cultured for an extended period of time in the presence of a specifically labelled ^{13}C -tracer which results in the incorporation of the tracer atoms into metabolic intermediates and products. The constraints

$$f_{\text{isotopomer}}(\vec{x}, \vec{v}) = \vec{0}$$

Is introduced to account for isotopomer balancing.

Chapter 3

Noise

3.1 Regulation of noise in the expression of a single gene

The aim of the paper was to investigate noise levels in gene expression for a single gene. Reporter system architectural design: a single-copy chromosomal gene with an inducible promoter. The goal is to ensure that only a single copy of the gene is present in the plasmid.

Scientists incorporated a single copy of the reporter, the green fluorescent protein gene (*gfp*), into the chromosome of *B. subtilis* easy to detect, quantifiable. They chose to integrate *gfp* into the chromosome itself, rather than in the form of plasmids, as variation in plasmid copy number can act as an additional and unwanted source of noise. Transcriptional efficiency was regulated by using

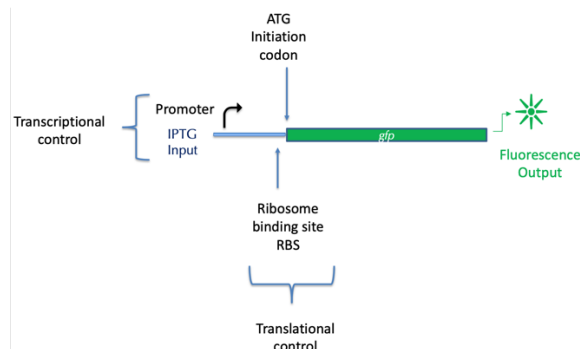


Figure 3.1: Reporter system architectural design

an isopropyl- β -D- thiogalactopyranoside (IPTG)–inducible promoter, Pspac, upstream of *gfp*, and varying the concentration of IPTG in the growth medium. Furthermore, translational efficiency was regulated by constructing a series of *B. subtilis* strains that contained point mutations in the ribosome binding site (RBS) and initiation codon of *gfp11*. The use of two different strategies to regulate transcriptional and translational processes introduces a potential bias in the relative contributions of these processes to biochemical noise. As a control, they constructed four additional strains whose transcription rates were altered by mutations in the promoter region of the reporter gene.

3.2. ADAPTIVE RESPONSE OF GENE NETWORK TO ENVIRONMENTAL CHANGES BY FITNESS-INDUCED ATTRACTOR SELECTION

Flow cytometry: scan cells one by one by scattering different lasers, detect fluorescence and register the results. It was seen that the phenotypic noise strength shows a strong positive correlation with translational efficiency, in contrast to the weak positive correlation observed for transcriptional efficiency.

Conclusions

- increased translational efficiency will strongly increase the variation in the expression of any naturally occurring gene.
- Phenotypic variation can be controlled by genetic parameters: low translation rates will lead to reduced fluctuations in protein concentration.

They hypothesize that several inefficiently translated regulatory genes have been naturally selected for their low-noise characteristics, even though efficient translation is energetically favorable. Therefore, depending on the metabolic pathway, some very important genes might be poorly translated for avoiding noise.

3.2 Adaptive response of gene network to environmental changes by fitness-induced attractor selection

Cells alter their gene expression in response to environmental changes or external signals to switch between coherent genetic programs in order to produce a phenotypic state, among many available, that best copes with the new environment. It is increasingly becoming clear that such genetic programs represent **attractor states**: discrete stable states of gene expression patterns generated by the dynamics of the regulatory interactions between the genes. The question is how cells switch into the appropriate attractor that is commensurate with the environmental condition. For instance, if the nutritional situation requires expression of gene A, how do cells switch into the attractor state in which A is stably expressed?

Attractor states are multiparameter complex systems that tend to attract sub-parameters that allow the system to persist in time. The existing paradigm is that cells have evolved a signal transduction machinery to sense the environmental change and transmit it to the gene regulatory network. There is an established link between the environment and the genes..

In the simplest case, such as in bacteria, the environmental signal may be a metabolite that directly regulates the transcriptional complex that controls the operon involved in its utilization (e.g. the lactose operon). In more complex systems, membrane receptor proteins sense environmental changes and trigger a cascade of intracellular molecular events involving “secondary messengers” or protein phosphorylation cascades that lead to concerted changes in the expression of several genes.

However, since the space of environmental conditions is much larger than that of cellular response programs, there is not a program for each condition, and cells need to choose the optimal program for a given condition. Therefore, it is unlikely that cells have evolved a specific signal transduction pathway for every environment it may encounter. In fact, infrequently occurring environmental conditions or unspecific, physical perturbations devoid of molecular specificity can evoke specific cellular programs, such as proliferation, quiescence or apoptosis.

3.2. ADAPTIVE RESPONSE OF GENE NETWORK TO ENVIRONMENTAL CHANGES BY FITNESS-INDUCED ATTRACTOR SELECTION

In order to study this phenomenon, the authors of the paper built a test system. The aim is to study adaptive responses to environmental changes without signal transduction machinery. In some conditions one of the two operons will be completely repressed/expressed, in other cases they will both be active (red and green fluorescence). E.g. turn on operon 1 for surviving glutamine lack. As a basic property of the circuit architecture, the cells harboring pALL7 can be in monostable and bistable behavioral regimes, depending on culture conditions. After several serial overnight cultures in Medium N, which does not impose restrictions on essential nutrients, the cells proliferated sufficiently fast so that the gene products of the two operons, including the repressors, were kept low due to dilution.

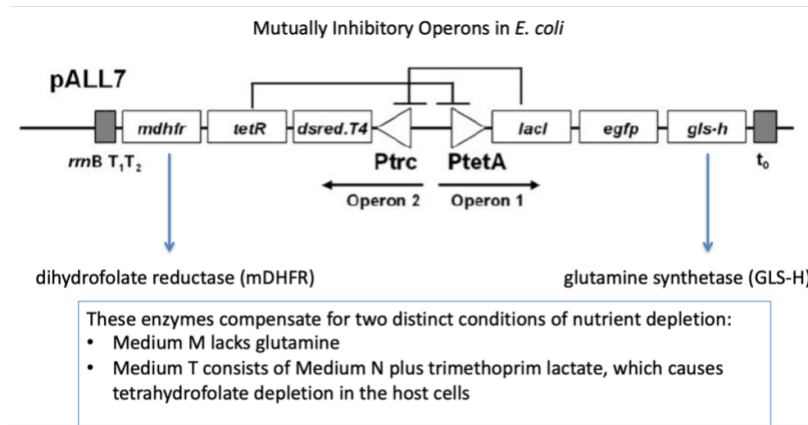


Figure 3.2: Test system

Consequently, mutual inhibition was too weak and did not produce bistability. In this monostable regime cells were reproducibly distributed around low levels of expression for both operons (blue dots). Two-color flow cytometry analysis indicate low levels of expression of Operon 1 (green fluorescence) and Operon 2 (red fluorescence) for each cell. Then, they changed the environmental condition by adding 170 ug/ml nalidixic acid reducing the specific growth rate by 30,40%. The balanced state of Attractor W is no longer stable. When either of the operons happens to express its encoded repressor at a slightly higher level than the other, the other operon is slightly suppressed, which in turn decreases the concentration of the repressor for the former operon, leading to a further increase in expression of the former.

Next they studied the “adaptive” response of the network to external changes by exposing the cells to culture conditions that require the presence of either enzyme (GLS-H or mDHFR) whose mutually exclusive expression is associated with the two attractors. Thus, asking whether cells can find the “adaptive attractor” that copes with the nutrient condition. For this purpose, they used two environmental conditions to implement the respective nutrient depletion: Medium M lacks glutamine, so that cells are required to synthesize it to keep up cellular activity. Cells carrying pALL7 can overcome glutamine depletion if glutamine synthetase (GLS-H) in Operon 1 is stably expressed, that is, when they are in Attractor 1. Conversely, Medium T consists of Medium N plus trimethoprim lactate, which causes tetrahydrofolate depletion in the host cells. In this case the host cells carrying pALL7 can overcome tetrahydrofolate depletion if mDHFR in Operon 2 is expressed, which is active when cells are in Attractor 2.

In the flow-cytometry measurements, they observed that each cell underwent a unidirectional shift

3.2. ADAPTIVE RESPONSE OF GENE NETWORK TO ENVIRONMENTAL CHANGES BY FITNESS-INDUCED ATTRACTOR SELECTION

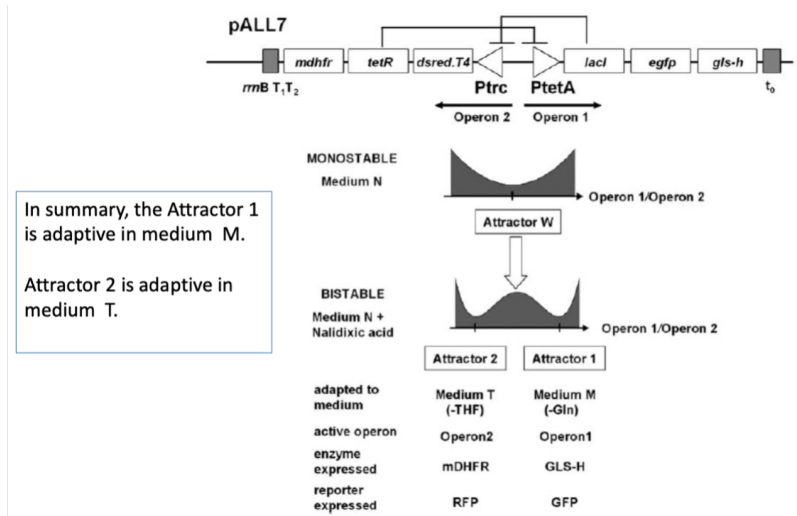


Figure 3.3: Attractors

in gene expression pattern toward the attractor expressing the adaptive enzyme in response to respective nutrient depletion. The nutrient depletions in medium M and T caused a selective, uni-directional shift towards the corresponding attractors.

Next they examined how cells adjust their gene expression adaptively to fluctuating environments. Cells carrying pALL7 were subjected to serial overnight culture with sequential medium changes in two different orders. Under good nutrient conditions both operons are on. It appears that cells without specific sensors are able to switch their gene networks in order to survive. What is the explanation of this?

1. an external signal that causes a commitment to the expression of a particular phenotype does so by somehow instructing the gene transcription apparatus to express the appropriate (set of) gene(s) in all cells, or
2. the signal merely promotes survival and expansion of the few cells that “happen” to express that desired phenotype

They found that the scenario 2 alone cannot explain the observed macroscopic shift toward adaptive attractor, but the scenario 1 indeed is necessary, as demonstrated by careful monitoring of the time course in which cells redistribute between each of attractors.

Wrong nutrient condition metabolism becomes inactive increase in noise (random turn on and off) survival. When the network state reaches an attractor that is adaptive, cells exhibit high cellular activity, increasing the turnover rates of mRNAs. This in turn suppresses the influence of gene expression noise. In contrast, for the non-adaptive attractor, which accordingly has low cellular activity, the metabolic rate is smaller, and hence, noise overwhelms the deterministic component of the dynamics of the network. This causes the cell to be kicked out of the non-adaptive attractor.

Due to its stochastic nature, attractor selection may not be as efficient as ordinary signal transduction, but it may prevent cells from dying in fluctuating environments. Attractor selection may

3.2. ADAPTIVE RESPONSE OF GENE NETWORK TO ENVIRONMENTAL CHANGES BY FITNESS-INDUCED ATTRACTOR SELECTION

facilitate the design of a network that can robustly respond in an adaptive manner to unknown environmental changes without requiring a large number of specific sensors and transducers. It may also be viewed as a sort of Darwinian preadaptation for the evolution of signal- specific transduction pathways when a particular new environmental condition becomes dominant and hence contributes to evolvability.

Chapter 4

CFPS and MAGE

4.1 Cell-free protein synthesis

Cell-free protein synthesis is also called IVVT (in vitro transcription translation). We take living cells, grow them up to a certain density, lyse the membrane to get the cytoplasm and use it to run a metabolic reaction. CFPS can be defined as the production of proteins using biological machinery without the use of living cells. The in-vitro protein synthesis environment is not constrained by a cell wall or homeostasis conditions necessary to maintain cell viability. CFPS enables direct access and control of the translation environment, which is advantageous for a number of applications including: optimization of protein production, optimization of protein complexes, study of protein synthesis, incorporating non-natural amino acids, high-throughput screens, and synthetic biology. For instance, incorporating non natural amino acids could be poisonous for the cell, but in this case this can be performed - as the cell is not viable anymore. George Curch is one of the pioneers in the field. He states the following advantages of CFPS:

1. rather than attempt to balance the tug-of-war between the cell's objectives and the engineer's objectives, in vitro biocatalysis focuses cellular resources toward an exclusive user-defined objectives.
2. cell viability constraints are removed.
3. transport barriers are removed, allowing easy substrate addition, product removal, system monitoring, and rapid sampling.

There are two main CFPS types:

- Cell extract: grow cell population, disrupt the membrane and isolate the metabolism hoping that the proteins are still viable. We have four major sources for extraction: Escherichia coli (ECE), rabbit reticulocytes (RRL), wheat germ (WGE), insect cells (ICE). We choose according to the kind of protein we are interested in, as protein production in E. coli and eukaryotes is quite different. All of these extracts are commercially available.
- The PURE system (E.coli only). Individual compotents for transcription machinery are isolated.

CFPS is quite important (remember tocopherol from Lecture 1). In 1960s it was not clear how to map DNA information to protein production. Nirenberg was the first to find the first codon sequence – UUU, phenylalanine.

4.1. CELL-FREE PROTEIN SYNTHESIS

4.1.1 Cell extract

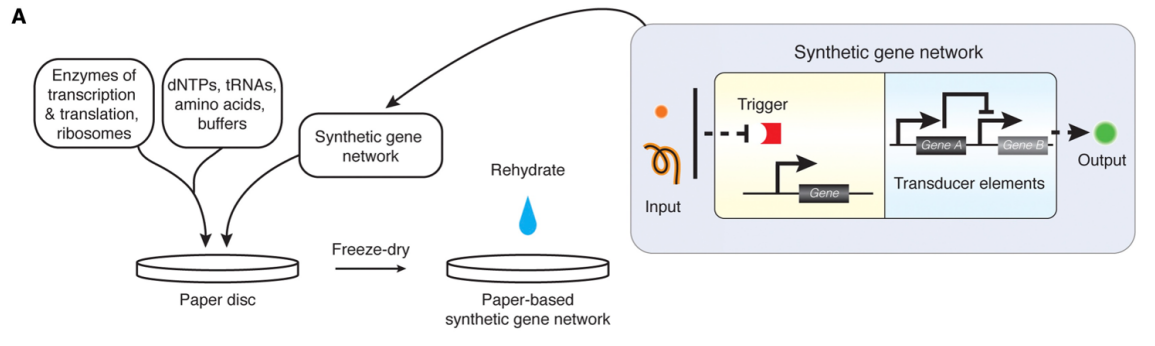


Figure 4.1: Cell-free protein synthesis procedure

The cell extract is put on a paper filter disc with basic components (dNTPs, tRNAs) and freeze-dried. The disk is rehydrated, hoping that functions are restored. If the synthetic gene network is active, we should get a tangible output. The same can be applied with plasmid DNA or inducers, we can check GFP expression. We get everything that the cell produces, also potential dangers e.g. proteases, lysosomes.

4.1.2 PURE system

The PURE system (Protein synthesis Using Recombinant Elements) is the reconstitution the *E. coli* (transcription) translation process in a test tube. It provides higher reaction controllability in comparison to crude cell-free protein- synthesis systems for translation studies and biotechnology applications. The PURE system stands out among translation methods in that it provides not only a simple and unique “reverse” purification method of separating the synthesized protein from reaction mixture, but also that the system can be tailor-made according to individual protein requirements.” Required materials are RNA polymerase, monomers, ribosome, tRNAs (aminoacylation), we require energy for reactions (ATP, GTP . . .). We also need all the enzymes e.g. kinases, phosphate economy control. The system needs to be buffered and protected from oxidation (DTT is used to avoid this),

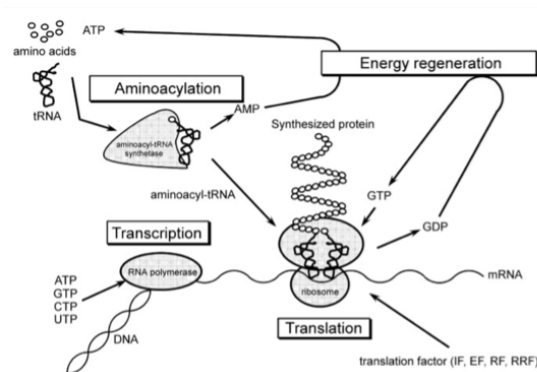


Figure 4.2: PURE procedure

pH control is required to avoid unwanted molecular interactions and protein folding, light should be also controlled. Everything has to be purified and reconstituted. It was possible to make with a good range of proteins; control is good, we can fully customize the system. The tradeoff is that we get a small protein yield with respect to other techniques. DNA is prepared for PCR with a FOR and REV primer + specific promoter. The ribosomes are supplied intact (avoid complex assembly).

Advantages over cell-free extract: because its components are defined, the PURE system does not contain some detrimental enzymes found in extracts. Individual components can be varied, added or subtracted depending in the application. PUREfrex Technology allows to obtain proteins for antibody generation, structural and kinetic studies and screening assays.

This system is good at making proteins from a gene, but not a replication (instead this occurs in a flow reactor). We have to keep feeding the system with the required materials, which are very expensive - not economically practical, everything is done in batch mode. Reducing a complex system will lead lacking the optimal yield, but ensuring good quality. The cell is maintaining concentrations of metabolites and proteins in a smart way, in this case instead we lose control. Glycosylation is an important process to take into consideration, involved in regulation.

4.2 Multiplex Automated Genome Engineering

The aim of MAGE is to improve changes in the genome (insertion/deletions) at specific positions. DNA replication occurs in section, the DNA is un-winded and the RNA polymerase starts from the leading strand (5' to 3' direction), then on the lagging strand from Okazaki fragments. We can overwhelm the system by adding artificial fragments, which should be similar to the wild type to create competition thermodynamics drives the process.

The efficiency of the MAGE process was characterized using a modified *E. coli* strain, mediated by a bacteriophage ssDNA-binding protein beta. The beta protein directs ssDNA to the lagging strand and promotes strand annealing. This strain also lacks mismatch repair. Sometimes the competition does not work, we have different outcomes. The oligos can be different e.g. ssDNA oligo, substitutions... The flanking regions must be faithful to ensure a correct recognition. Linear DNA molecules are flanked by homologous sequences (40-50 bp are more efficient) to target DNA sequences (1-60 nt). It is possible to insert N variations in the case in which it is not sure which modification to make. Each site can have variation and its own pool of oligos, massively parallel gene editing process.

Process: grow cells until a certain density is reached, include ss-oligos and obtain replacement by electroporation. Some cells do not survive electroporation, some do not take up external DNA, some take some oligos and get genetically modified. After some recovery time, cells are grown again and the process is repeated. At any time, it is possible to harvest some cells for screening, selection or genotyping.

For each cycle, a certain percentage of population is genetically modified. To overcome low efficiency, the oligo-recombineering protocol is iterated on the same cell population over multiple cycles using the same oligo species. In this fashion, the population is enriched for mutants containing the desired sequence conversions. Each full cycle takes 2-3 h, depending on the growth rate of the cells.

The relative abundance of mutants in the population M can be approximated by $M = 1 - (1 - RE)^N$

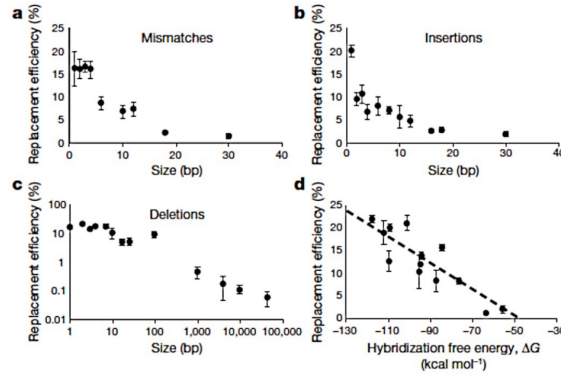


Figure 4.3: Replacement efficiency in mismatches, insertions, deletions

where N is the number of cycles and RE is the allelic replacement efficiency per cycle. RE is highly dependent on the type of target conversion (mismatch, insertion, deletion) and the size of the conversion. The efficiency depends on the amount of genetic modification we wish to add (figure), as it becomes increasingly difficult for fragments to compete when they diverge from the original sequence.

Genetic screens can be in the form of direct genotypic methods such as PCR or DNA sequencing, or phenotypic screening or selection methods such as colorimetry, growth rate, or antibiotic resistance. By computing $N = \frac{\log(1 - F)}{\log(1 - RE)}$, we can find the number of cycles N needed to produce mutation size of b base-pairs at a frequency of at least F in the population. For example, the number of cycles needed to generate mutants with a 6 bp chromosomal mismatch to a frequency of 0.25 (i.e., 25%) in the population with an oligo folding energy of -5.4 kcal/mol (predicted through MFold; Markham and Zuker, 2005) is $N = \log(1 - 0.25) / \log(1 - 0.26e^{-0.135 \times 5.4}) = 2.0$ cycles, and to a frequency of 0.50 (i.e., 50%) is $N = 4.9$ cycles. Thus, one would expect from a PCR screen that at least one in four cells would show conversion after two cycles and one in two would show conversion after five cycles of oligo-recombineering. This frequency is high enough that alterations can now be made without selection. With optimized protocols, over 50% of the cells that survive electroporation contain the desired change.

Multiple cyclings with multiple targets can lead to a combinatorial explosion with the main limitation being the cell population size (around 10^9 cells). MAGE test system: the targeted lacZ region was sequenced in 96 random clonal isolates after MAGE cycles 2, 5, 10 and 15 that provided a snapshot of the genotypic variation in each population. We have consecutive N30 oligo, interspersed n6 oligo and consecutive N6. While cycles increase, we see an increase in mutations. The depth at which MAGE generates diversity is determined by a combination of three factors:

1. the degree of sequence variation desired at each locus;
2. the number of loci targeted;
3. the number of MAGE cycles performed.

4.2. MULTIPLEX AUTOMATED GENOME ENGINEERING

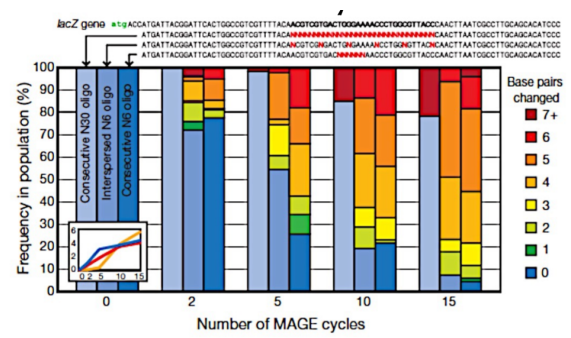


Figure 4.4: Sequence diversity generated across three separate cell populations as a function of the number of MAGE cycles

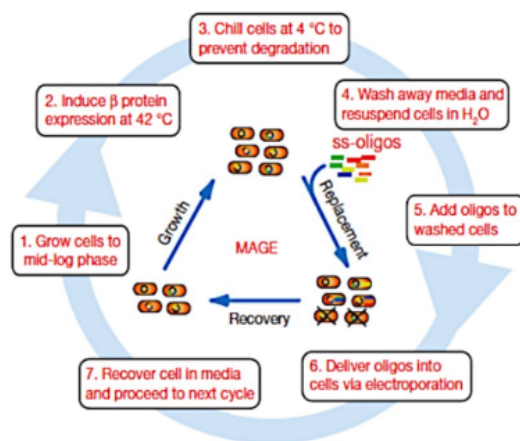


Figure 4.5: MAGE automation

4.2.1 MAGE automation

To demonstrate an application of the MAGE process, they optimized metabolic flux through a biosynthesis pathway to overproduce the isoprenoid, lycopene, in an *E. coli* strain (EchW2) that contained the pAC-LYC plasmid that is necessary for the final steps of lycopene production. Specifically, for each of the 20 genes, 90-mer oligos containing degenerate ribosome binding site (RBS) sequences (DDRRRRRRDDDD; D = A, G, T; R = A, G) flanked by homologous regions on each side were used, with a total pool complexity of 4.7×10^5 . Additionally, four genes (ytjC, fdhF, aceE, gdhA) from secondary pathways were targeted for inactivation by oligos that introduced two nonsense mutations in the open reading frame, further improving flux through the DXP pathway. Therefore, they optimized 24 genes simultaneously to maximize lycopene production. As many as 15 billion genetic variants (4.3×10^8 bp variations per cycle for 35 MAGE cycles) were generated and screened (intense red pigmentation on Luria-Bertani agar plates). Variants were isolated from 105 colonies screened after 5–35 cycles of MAGE and sequenced. We have a huge variation in growth rate depending on the genetic background. We would expect a positive correlation between the mutation rate and growth rate, but this is not straightforward. For lycopene production in an *E. coli* strain, a 5 fold improvement in yield after three days was observed (pretty fast).

On balance, MAGE is fast and efficient tuning of genetic diversity in *E. coli*. It may be applied to many MGE outcomes and may be applicable to other organisms.

4.3 CFPS and MAGE

Goal: to test the Multiplex Automated Genome Engineering (MAGE) strategy to simultaneously modify and co-purify large protein complexes and pathways from the model organism *Escherichia coli* to reconstitute functional synthetic proteomes in vitro. Ni-NTA Agarose is an affinity chromatography matrix for purifying recombinant proteins carrying a His tag. By adding 6 His tags on the C or N terminus on the protein of interest, we can isolate the full complement for the protein through affinity chromatography.

In the study, nine total strains were constructed for the ensemble PURE system. Insertion of His-tag sequences into all target genes in the ePURE strains was characterized by PCR and subsequently verified by sequencing. By application of over 110 MAGE cycles, they successfully inserted hexa-histidine sequences into 38 essential genes in vivo that encode for the entire translation machinery. The colour codes are from four different population of cells. With this amount of changes, nothing managed to grow in the first population. It was required to split into different sets, as the overall metabolic burden is too much.