A dynamic genome-scale modelling of *E. Coli* fermentation (dGSMEF)

Lulu Zheng

2017-03-08

1. Introduction

Escherichia coli is an important host organism in industry for the production of many biopharmaceutical proteins. Therefore, it is crucial for us to improve the efficiency of fermentations to achieve high product yields. *In silico*, constraint-based genome-scale modelling with flux balance analysis (FBA) has become a popular method in metabolic engineering for fermentation performance improvement. However, this classic FBA method can only be used to study the static metabolic states of cells. In fermentations, environments are always changing and cell growth and intracellular metabolism are also highly dynamic. This drives us to develop a novel approach, which combines the key advantages of FBA with dynamic simulation of the extracellular environment. The aim of this project is to develop such a dynamic FBA approach able to describe the whole behaviour of fermentation processes. Similar to FBA, the key assumption that must hold for this approach is that pseudo steady-state is maintained for intracellular metabolism.

Based on the genomic gene annotations, a genome-scale model which contains all known metabolic stoichiometric reactions for our strain was firstly constructed. And then, this model was analyzed in a dynamic framework, which enables an interaction between the metabolism and changing extracellular environments, to simulate substrate, biomass, and recombinant protein concentrations with time for growth in batch or fed-batch cultures.

Here we present a dynamic model that integrates the metabolic network with the dynamics of cellular metabolites, biomass composition, and recombinant protein synthesis constraint for *E. coli* fermentation. The model consists of two major parts: (1) a flux-balance based genome-scale model with 1973 metabolites, 2784 reactions and 1350 genes and (2) 393 process variables of interest that describe biomass, metabolites and protein concentration profiles with time. The model incorporates growth-rate dependent biomass composition and energy requirements, maintenance energy theory for the host and recombinant cell, proteome resource allocation model to balance the bi-objective optimization of maximum growth and recombinant protein synthesis, as well as dynamic flux balancing. The parameters of the model were estimated using experimental data. Using the model, we are able to predict the behaviour of our strain producing proteins with many conditions, such as rich medium in batch phase, acetate secretion and re-utilization after depletion of glucose, glucose minimal medium with/without complex medium yeast extract (YE), etc.

In summary, the presented approach represents a powerful tool for *E. coli* production process simulation and ultimately optimization. Both intracellular metabolic fluxes and fermentation process could be monitored. Therefore, this dynamic model can be applied to optimize medium and bioprocess for improving the recombinant protein production.

This manual will guide us how to run this model step by step in the following parts.

2. Model workflow

The proposed model is composed by six blocks as depicted in the scheme shown in Figure 1. The first block, genome-scale model reconstruction, establishes the foundation of modelling of cellular metabolism. And the second block, model parameters determination experimentally, defines the constraints, such as maximal glucose uptake rate, to the genome-scale model. With this constaint-based genome-scale model, FBA approach has been shown to provide meaningful predictions in many organisms at a particular steady state. However, this static FBA can only identify cellular metabolic fluxes, and cannot provide any information on

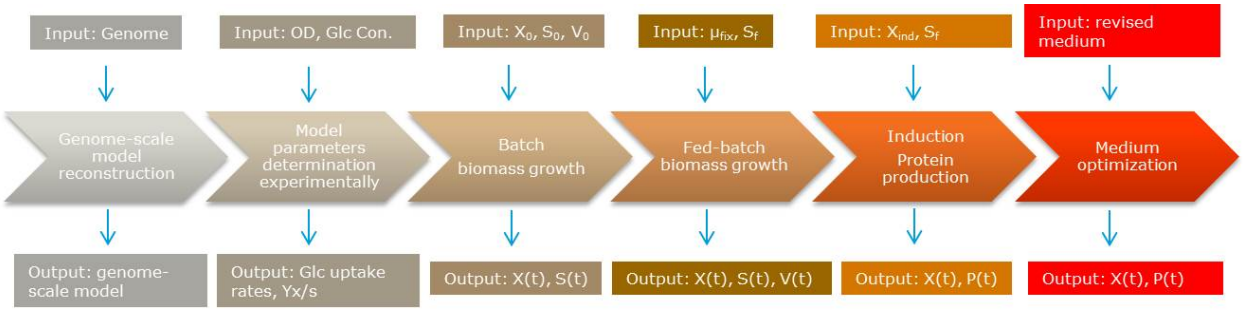


Figure 1: Model blocks and variables. For interpretation of variables, please refer to Table 1

the metabolite concentrations, as well as on the dynamic characteristics of these metabolic fluxes. Therefore, in a changing enviroment such as fermentaion process, we need to modify FBA, where the flux balance analysis problem is solved along with constraints on the change of metabolite levels at specific time instants. The followed three blocks, covering batch, fed-batch, and induction cultures, calculate the dynamics of external metabolites, biomass and recombinant protein concentrations. Differential Equations are listed in Table 1. We extend FBA, known as dynamic FBA, to account for this dynamics. The dynamic FBA is implemented with static optimization approach (SOA). First, the simulation time is divided into small periods which are assumed to be in quasi-steady state; Second, for every time step, an FBA problem is solved and the fluxes are integrated over the time period and extracellular concentrations are calculated, accordingly. Furthermore, the flux optimizaiton is achieved by parsimonius FBA (pFBA), which minimize all fluxes within the solution space of the growth optimum, to minimize the requirement for enzyme expression. In particular, in induction protein production block, when the culture reaches the desired biomass concentration (X_{ind}), the FBA problem faces a new challenge, that is, the bi-objective optimization of both maximum growth and recombinant protein synthesis rates. To deal with it, a novel proteome resource allocation theory is applied. According to this theory, the growth reduction induced by heterologous protein expression is a simple consequence of proteome allocations. And more heterologous proteins, less growth rate. The final block, acctually, is one of our model applications to the medium design. By comparing the whole behaviour and interal metabolism, we may obtain some helpful insights to guide us improve the quantity and/or quality of protein expression.

Table 1. Ordinary differential equations and variales nomenclature

Equations	Nomenclature	Description	Nomenclature	Description
$F_s = \frac{1}{S_f} qXV$	F_s	feeding flow rate (L·h-1)	X_0	initial batch biomass concentration (gDCW·L-1)
	S_f	glucose concentration in the feeding solution (mmol·L-1)	S_0	initial batch glucose concentration (g·L-1)
$D = \frac{F_s}{V}$	q	specific glucose efflux rate determined by FBA (mmol·gDCW-1·h-1)	$Y_{X/S}$	biomass yield (gDW·g-1 glucose)
$\frac{dV}{dt} = DV$	X	biomass concentration (gDCW·h-1)	V_0	initial volume (L)
	V	total volume (L)	μ_{fix}	Fixed specific growth rate during fed-batch (h-1)
$\frac{dX}{dt} = (\mu - D)X$	D	dilution rate (h-1)	X_{ind}	biomass concentration at induction (gDCW·L-1)
	μ	specific growth rate (h-1)	P	specific protein in mass units (g·L-1)
$\frac{dS}{dt} = qX + D(S_f - S)$	S	glucose concentration (mmol·L-1)		
$\frac{dS'}{dt} = q'X - D*S'$	S'	other substrates concentration (mmol·L-1)		
	q'	specific efflux rate determined by FBA (mmol·gDCW-1·h-1)		

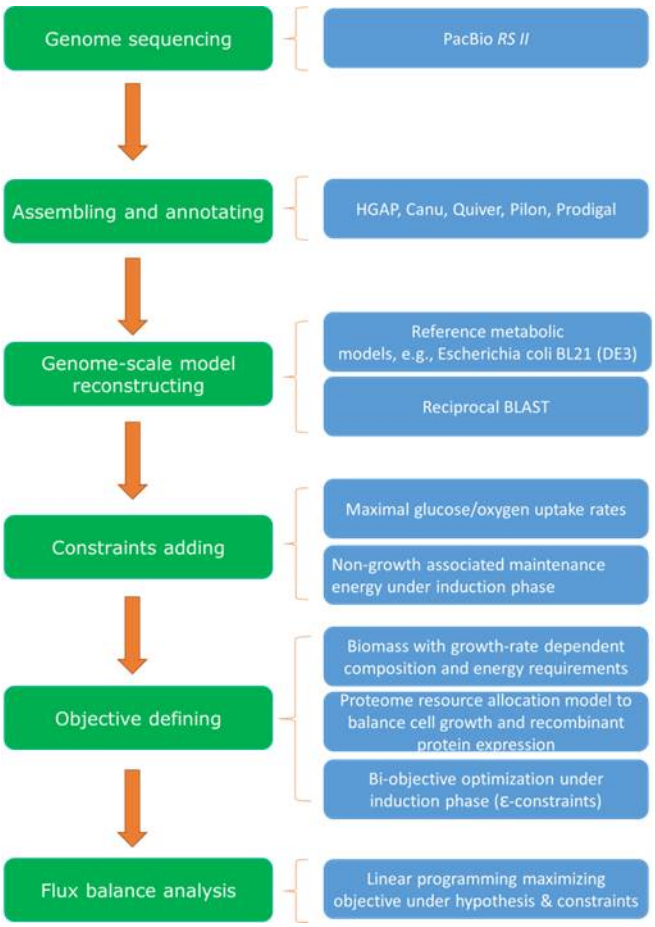


Figure 2: The genome-scale model reconstruction process

2.1 Construction of a constraint-based genome-scale model

The central component of the FBA is a stoichiometric matrix, which is converted from a genome-scale metabolic network model. Figure 2 shows the reconstruction process of this genome-scale model.

2.1.1 *De novo* sequencing and assembly of HNC47 by PacBio *RS II*

Single molecule, real-time sequencing by PacBio has become the gold standard for finishing microbial genomes. Thus, we sequenced HNC47 genomic DNA by PacBio RS II of Tianjin Pangu Gene company, with a single library (<20kb). Sequencing statistics data shows it generates 1237Mb data, containing 86228 raw reads (polymerase reads), of which the average length and quality are 14.346kb and 0.84, respectively. After consensus calling, the average length and quality of circular consensus reads (reads of insert) are 8.682kb and 0.87, respectively. The QC-loading report shows that the percentage of ZMWs that are productive and sequencing is more than 50%, demonstrating the well-loading of this sequencing experiment.

Considering the sequencing coverage is more than 200X, only PacBio data is supposed to be enough to assemble a fully complete genome. In the next steps, the assembly was only done based on the PacBio data. Previous illumina data for HNC47 was just used for final polishing. Raw reads are firstly split into longest "seed" reads and shorter reads by using a length threshold. This threshold is typically chosen such that we have at least 30X of coverage in the "seed" group. Here, I assume HNC47 genome size is 4.5M, which will be used as a "GenomeSize" parameter value in all three software. The shorter reads are then used to correct the seed reads through alignment and consensus resulting in corrected reads. These could be done

by HGAP, Canu or Celera Assembler. These corrected long reads have high enough quality for assembly. After we obtain the genome sequence, a contig polishing software called Quiver is used to further improve the accuracy of the assembly. For HGAP, it has been integrated into PacBio's SMRT analysis package, which could directly start from the original PacBio bas/bax.h5 files. After several steps including filtering raw reads, preassembly, assembly and consensus polishing, it could generate final contig(s). In contrast, Canu and Celera Assembler began with raw subreads in FASTQ format. And then, pbalign with blasr algorithm was applied to map raw reads against draft assemblies. Quiver was lastly used to help derive a highly accurate consensus for the final assembly. Obviously, Celera Assembler got a much shorter contig, compared to HGAP and Canu. Further comparisons showed that this contig missed some genes, which are found to be existing by illumina sequencing results. Therefore, only HGAP and Canu assembly contigs are remained. Here, Canu also got a very short contig, which is almost composed of T and only a few reads supported it. This contig was discarded.

But the work was not finished. Dotplots against BL21(DE3) and self both showed that the genome should be a circular molecule, the beginning and end of the contig containing the same sequence. Naturally, this genomic circularity is true for E.coli. It results in reads mapping to both locations during the polishing step, as such the reads have a low mapping score, and are not used to call consensus. In this case, a polished assembly will have low quality in these regions. To address this problem and generate a blunt ended circular sequence that has high quality consensus throughput, Amos was then used. After selecting a new starting position (actually it is the same as BL21(DE3)), the beginning and end were joined, and a new contig was generated. Another quiver polishing was done again. By HGAP, this results in a contig with length of 4,487,376 bp, QV $>$ 60 (99.9999%), no overlap between the beginning and end, mapped reads/post-filtered reads ratio exceeding 90%, and even coverage distribution across the genome. All these metrics showed that this final assembly performance has a high level, and could be taken as HNC47 final reference genome. It should be noted, now HGAP and Canu generated nearly the same contig, except 38 single nucleotide indels, where this may result from the intrinsic weakness of PacBio platform.

Since we also have illumina paird-end sequencing data of HNC47 genomic DNA, this final reference genome could be further improved by using short read mapping, in particular small indels. BWA mapped the illumina paird-end data to HNC47 reference genome, and then Pilon removed 30 remaining indels in this ~4.5M genome despite Quiver calling $>$ 60. A Quiver re-polishing for correction was done, generating the final HNC reference genome with a length of 4,487,375 bp.

2.1.2 Genomic annotation

The annotation includes the following contents: 1) Protein-coding genes and corresponding coding sequence (CDS); 2) ribosomal RNA genes (rRNA); 3) transfer RNA genes; 4) non-coding RNA genes; 5) CRISPR.

Prokka combined with other tools including Prodigal, Barrnap, Aragorn, Infernal and MINCED is used to annotate this prokaryote genome. Result files with various formats are deposited in "R:\ERC\NNRCC_ERC\ERCPE\Database\HNC47".

2.1.3 Reconstruction of genome-scale model

Others' study has shown that starting from the highly curated content of a closely related organism produces a more accurate model than the currently available automated methods. Therefore, gene sequences from Escherichia coli BL21(DE3), B str. REL606, K-12 MG1655, and W are used for identifying orthologs. Reciprocal blastp method is applied to identify best bidirectional hits and an e-value of 1e-10 cutoff is used for assigning orthologs. 3992 genes (96.78%) have orthologs. Remaining genes that are missing orthologs in the original models are deleted from the model for the HNC47 strain. Take BL21(DE3) as a basis, additional reaction content is added from B str. REL606, K-12 MG1655, and W. Compared to BL21(DE3), it misses 13 reactions, and adds 1 new reaction. Most missing reactions won't affect the model's growth, except the reaction of Alanineracemase, which belongs to the "Alanine and Aspartate Metabolism" subsystem and is catalyzed by gene dadX or alr. Considering dadX is cloned into the plasmid to maintain HNC47 strain's normal growth, dadX is recovered and the reaction is also remained.

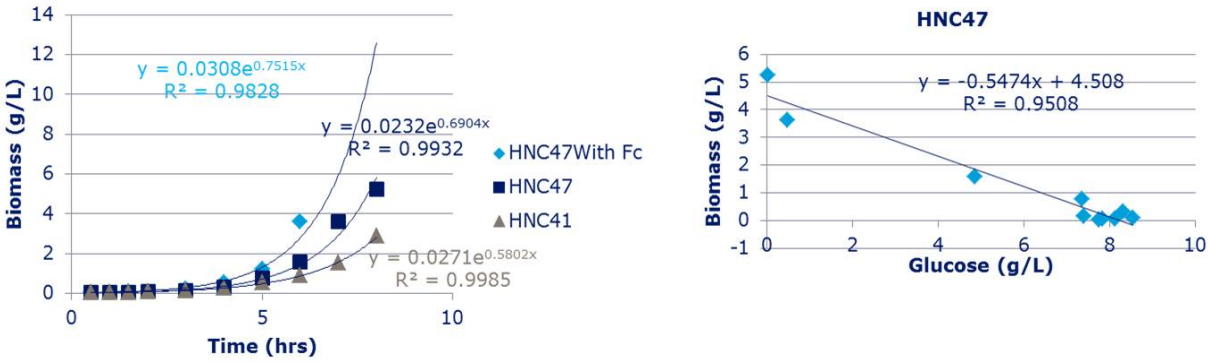


Figure 3: HNC47 maximum aerobic glucose uptake rate, determined from batch experiments as the growth rate (0.69 h-1 HNC47) divided by the biomass yield (0.547g DW/g glucose)

(Co-)Orthologs between HNC47, BL21(DE3), B str. REL606, K-12 strain MG1655 and W strain are identified by Reciprocal BLAST (Bidirectional Best Hit) based Proteinortho with default parameters. The GEM is reconstructed based on these ortholog genes. For genes s0001, gltL, fucA, wbbJ, fucP, xapA, ompC, tynA, yedL, rbsD, vsr, lsrD and fliZ, they're not mapped to a genome annotation of BL21(DE3). So, in the building of our HNC47 GEM, the corresponding reactions are maintained. Considering dadX is cloned into the plasmid to maintain HNC47 strain's normal growth, dadX is recovered and the reaction is also remained. Table 2 outlines the statistics of GEMs. Likewise, the major difference of new HNC47 GEM is the introduction of ybhE gene insertion, when compared to BL21(DE3). The new adding reaction aided by ybhE gene introduces an important process in the Pentose Phosphate Pathway, which promotes the generation of NADPH, pentoses as well as Ribose 5-phosphate, a precursor for the synthesis of nucleotides. Eventually, HNC47 metabolic model has 1350 genes involved in 2784 reactions with 1973 metabolites, as Table 2 shows. GSM files could be retrieved from "R:\ERC\NNRCC_ERC\ERCPE\Database\HNC47\Genome-scale model". SBML and JSON formats are provided. And in the R sybil package, it is suggested to use SBML format without fbc.

Table 2. The statistics of E.coli GEMs

Component (#)	BL21(DE3) AM946981.2	B str. REL606 CP000819.1	HNC47	HNC41
Metabolites	1945	1953	1973	1942
Reactions	2742	2749	2784	2728
Genes	1337	1329	1350[1]	1330

2.2 Model parameters determination experimentally

In order to apply the model we first needed to determine strain-specific parameters, in particular glucose uptake rate. The fully specified model was then used to compare predictions with experimental data.

The model requires maximum glucose uptake rate specified for metabolic simulations of batch and fed-batch cultures to restrict the metabolic capacity for glucose utilization. It's defined as the ratio of the growth rate to biomass yield in batch experiments (Figure 3). For aerobic batch cultures, we have determined this value to be 7 mmol/gDW/h for HNC47 and for other strains (Table 3). Maximal oxygen uptake rate in Table 3 is identified by FBA as the value that would give the correct experimentally determined growth rate when using the core biomass reaction. It seems surprised that the growth rate of HNC47 with Fc is a little higher than HNC47 without Fc plasmid. The reason still remains unclear.

[1]One additional reaction is also added to the model, when the induction phase begins. It is the recombinant protein synthesis reaction according to its amino acid sequence. The energy cost of protein synthesis is assumed to be 4.306 mmol ATP/mmol amino acids.

Table 3. Experimental determination of model parameters with batch cultivation on FDM + Glucose (10g/L)

Strain	Growth Rate (/h)	Glucose uptake rate (mmol/g DW/h)	Biomass yield rate (g DW/g glucose)	Oxygen uptake rate (mmol/g DW/h)
HNC41 (PGL-)	0.58	7.7	0.418	11
HNC47 (PGL+)	0.69	7	0.547	11
HNC47 with Fc	0.76	9.58	0.4413	11.5

2.3 Batch biomass growth model

FBA can be used to examine dynamic processes such as microbial growth in batch cultures by combining FBA with an iterative approach based on a quasi-steady-state assumption (static optimization-based dynamic FBA). The substrate concentration (Sc) (mmol/L) is determined from the substrate concentration predicted for the previous step (Sco) or from the initial substrate concentration if it is the first time step:

$$S_c = S_{co}$$

The substrate concentration is scaled to define the amount of substrate available per unit of biomass per unit of time (mmol gDW-1h-1):

$$Substrateavailable = \frac{S_c}{X\Delta(t)}$$

where X is the current cell density and Xo is the cell density from the previous step. FBA is then used to calculate the substrate uptake (Su) and the growth rate (Âµ).Concentrations for the next time step are calculated from the standard differential equations:

$$\frac{dX}{dt} = \mu(X) \rightarrow X = X_0 e^{\mu\Delta(t)}$$

2.4 Fed-batch biomass growth model

In contrast to the batch mode, during the fed-batch mode, one or more nutrients (substrates) are fed (supplied) to the bioreactor and in which the product(s) remain in the bioreactor until the end of the run. When the initial glucose is consumed out and also the dissolved oxygen consumption is reduced, batch phase ends and glucose feeding will start. Feeding rate is performed according to a predefined exponential feeding profile based on mass balances and substrate uptake, keeping the specific growth rate at a fixed value.

Specifically, if S represents the substrate concentration (mmol/L) (except substrates such as glucose fed in the feeding stream), q the specific efflux rate determined by FBA (mmol/gDCW/h), X the biomass concentration (gDCW/h), D the dilution rate (h-1), then we could have the following equation according to mass-balance law:

$$dS/dt = qX - DS$$

2.5 Induction protein production model

Induction phase starts when the IPTG pulse addition is done, and then, recombinant protein starts to express. We first add this recombiant protein expression reaction into our genome-scale model, according to precursor balances and energetic requirements:

(6) ala___L[c] + (6) arg___L[c] + (11) asn___L[c] + (9) asp___L[c] + (6) cys___L[c] + (11) gln___L[c] + (17) glu___L[c] + (11) gly[c] + (5) his___L[c] + (4) ile___L[c] + (17) leu___L[c] + (17) lys___L[c] + (3) met___L[c] + (9) phe___L[c] + (19) pro___L[c] + (24) ser___L[c] + (14) thr___L[c] + (8) tyr___L[c] + (23) val___L[c] + (4) trp___L[c] + (964.544) atp[c] –> (1) Protein[c] + (964.544) adp[c] + (964.544) pi[c]

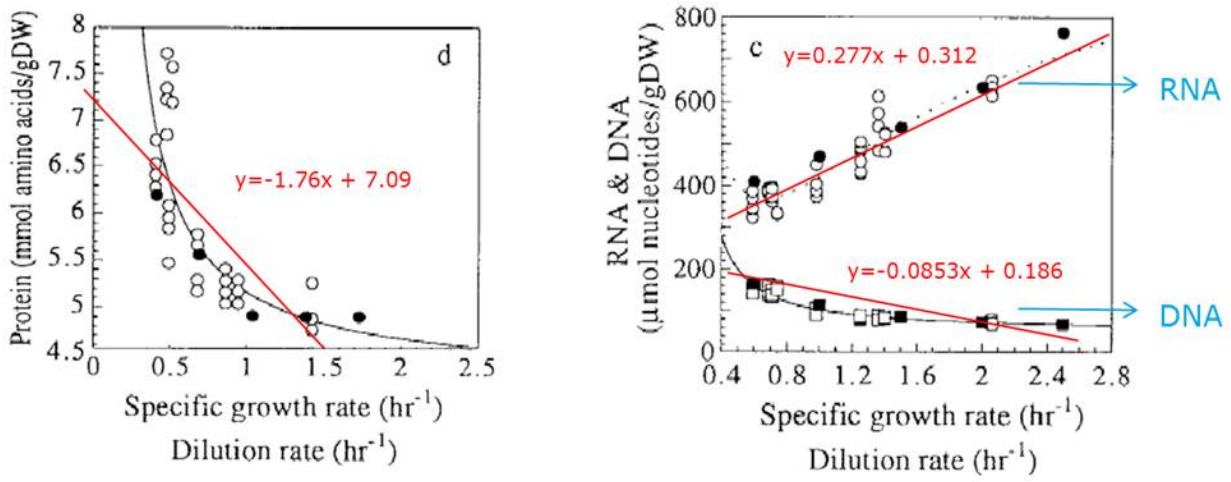


Figure 4: Variation in macromolecular composition of E. coli with growth rate, and the corresponding linearization between macromolecular composition and growth rate. All data are referred to J. Pramanik, J. D. Keasling. BIOTECHNOLOGY AND BIOENGINEERING, 1997

2.5.1 Growth-rate dependent biomass composition

In the constraint-based metabolic flux analysis, a prominent feature is that it is always required to define a reasonable objective function. And biomass objective function that describes the growth requirements of a cell is usually selected, e.g., in our batch and fed-batch cultures. The requirements contain the cell's chemical composition, i.e., protein, RNA, DNA, lipids, and cofactor content, as well as energy. Concretely, an average E. coli B/r cell growing exponentially at 37 C under aerobic conditions in glucose minimal medium with a doubling time of approximately 40 min has a dry weight of 2.8×10^{-13} g. The dry weight is 55% protein, 20.5% RNA, 3.1% DNA, 9.1% lipids, 3.4% lipopolysaccharides, 2.5% peptidoglycan, 2.5% glycogen, 0.4% polyamines, and 3.5% other metabolites, cofactors, and ions (Neidhardt, 1987). The types and amounts of precursors required to synthesize these macromolecules at a given growth rate could also be determined from the composition of each of the macromolecules: the amino acid composition of proteins, the nucleotide composition of RNA and DNA, the phospholipid composition, and the fatty acid composition from experimental data or primary literatures. However, the macromolecular composition is not the same for cells growing at different rates. Figure 4 shows the macromolecular composition of the cell changes with growth rate. Here, I modified their original fits describing the macromolecular composition such that all fits were linear:

$$Biomass\ content\ of\ a\ given\ molecular\ species\ (mmol/gdw) = slope \times \mu + intercept$$

In practice, when calculating the macromolecular content of the cell, the average u (growth rate) over the previous 2h was calculated first before computing the resulting biomass composition. In addition, the amino acid composition of proteins, the nucleotide composition of RNA and DNA remain the same all the time

2.5.2 Growth associated ATP maintenance (GAM)

The GAM reaction accounts for the energy (in the form of ATP) necessary to replicate a cell, e.g., for macromolecular synthesis (e.g., proteins, DNA and RNA). When experimental data is not available, the GAM can be estimated by determining the energy required for macromolecular synthesis. Therefore, the total amount of macromolecule (protein, DNA and RNA) is determined from databases or other resources. For example, average 4.306, 0.4 and 1.372 of phosphate bonds are necessary to synthesize a protein, RNA and DNA molecule, respectively. These phosphate bonds are accounted for by adding ATP hydrolysis to the biomass reaction ($xATP + xH2O \rightarrow xADP + xPi + xH+$, where x is the number of required phosphate

bonds). Note that this estimate will be too low, as other growth-associated cellular processes also require ATP. As the macromolecular composition of the cell changes with growth rate, so must the energy requirements to synthesize these macromolecules, which were correlated with the macromolecular needs. Because protein is one of the most energetically expensive macromolecules and because the relative amount of protein decreases with increasing growth rate, the total energy expended by the cell (per g DW) actually decreases with growth rate.

2.5.3 Non-growth associated ATP maintenance (NGAM)

In the constraint-based metabolic models, it also includes an ATP hydrolysis reaction ($1ATP + 1H2O \rightarrow 1ADP + 1Pi + 1H+$), which represents NGAM requirements of the cell to maintain, e.g., turgor pressure. However, under different conditions that the cell faces different pressures, the value for this reaction rate should not be constant. For instance, for recombinant cells, additional expression of recombinant proteins would bring metabolic burden to the cell, which would increase the NGAM requirement of the cell. Therefore, it's important to take this factor into account in our case. Pirt proposed a theory of constant maintenance energy in 1975, that is:

$$q = \frac{\mu}{Y_G} + m1,$$

where q is the specific rate of utilization of the energy source (glucose uptake rate in the model), μ is the specific growth rate, Y_G is the maximal growth yield and m1 is the constant maintenance energy coefficient, which is independent of the growth rate. Moreover, by putting $q = \frac{\mu}{Y}$ where Y is the actual growth yield, we could obtain:

$$\frac{1}{Y} = \frac{1}{Y_G} + \frac{m1}{\mu}$$

By computing the maintenance coefficient m1 for the host before induction and recombinant cell after induction, we are able to know the increased NGAM requirement due to the metabolic burden during the induction phase. Table 4 lists the biomass and glucose time-course data during the MIC-1 expression. Based on the above Pirt's theory, m1 for the recombinant cell with MIC-1 is 0.2459 g Glc/gDW/h (Figure 5). To convert m1 to NGAM, the following codes could be run:

Table 4. Biomass and glucose data during the MIC-1 expression

Time after induction (hr)	DCW (g/L)	Volume (L)	Delta DCW (g)	Delta Glucose (g)	u (h-1)	1/u (h)	1/Yx/s
0	21	0.107					
1	27	0.11134	0.75918	2.17	0.291074	3.43555	2.858347
2	34	0.11568	0.92694	2.17	0.268763	3.720753	2.341036
3	39	0.12002	0.74766	2.17	0.174032	5.746078	2.902389
4	42	0.12436	0.54234	2.17	0.10963	9.121577	4.00118

```
# Convert m1 to non-growth associated maintenance (NGAM) energy
library(sybil)
library(sybilSBML)
library(glpkAPI)
model = readSBMLmod("../data/HNC47_without_fbc.xml")
```

```
## reading SBML file ...
## OK
## getting the model ...
## OK
```

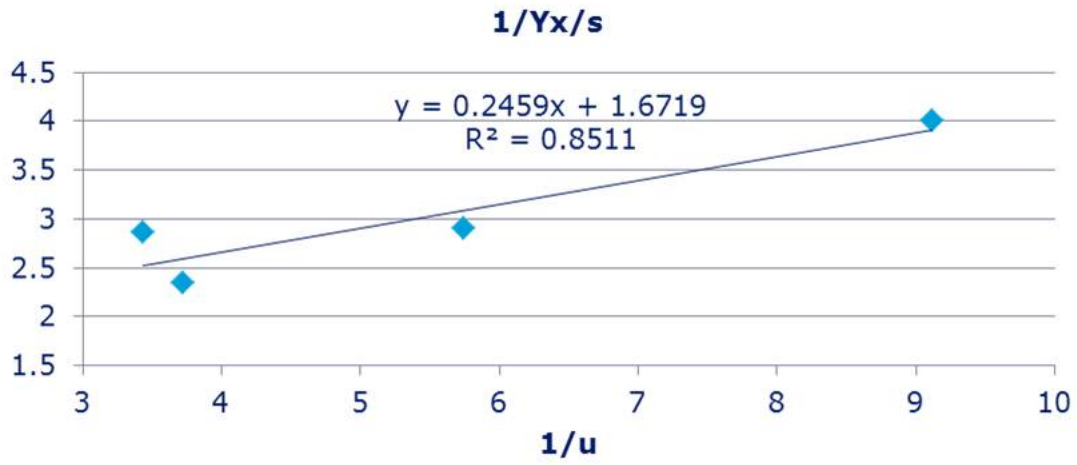


Figure 5: Maintenance coefficient for the recombinant E. coli growing on glucose as the limiting nutrient

```
## creating S and parsing constraints ...
## OK
## GPR mapping ...
## OK
## cleaning up ...
## OK
## validating object ...
## OK
```

```
lowbnd(model)[react_id(model) == 'EX_glc__D(e)'] = - 0.2459/180.16*1000 ## -1.364898 mmol Glc/gDW/h
model = changeObjFunc(model,c('BIOMASS_Ec_iJO1366_core_53p95M', 'ATPM'),c(0,1))
sol = sybil::optimizeProb(model,algorithm="fba",retOptSol=TRUE);
u_max = lp_obj(sol);
u_max
```

```
## [1] 32.0751
```

```
# reset the model
lowbnd(model)[react_id(model) == 'EX_glc__D(e)'] = - 20
model = changeObjFunc(model,c('BIOMASS_Ec_iJO1366_core_53p95M', 'ATPM'),c(1,0))
```

It outputs 32.075 mmol ATP/gDW/h, which is NGAM with MIC-1 case. Accordingly, for Fc under 50% glucose and 25% glucose + 10% YE feeding conditions, m1 and NGAM are 0.0806 g Glc/gDW/h and 10.51 mmol ATP/gDW/h, 0.0486 g Glc/gDW/h and 6.34 mmol ATP/gDW/h, respectively. And meanwhile, this m1 is supposed not to include the energy cost of recombinant protein synthesis, because μ actually represents the overall characteristics of the cell with recombinant protein expression (without recombinant expression, it's foreseeable to observe a larger μ). Obviously, the metabolic burden caused by MIC-1 is much larger than that by Fc.

2.5.4 A new proteome allocation model to balance resources distribution between biomass and recombinant protein

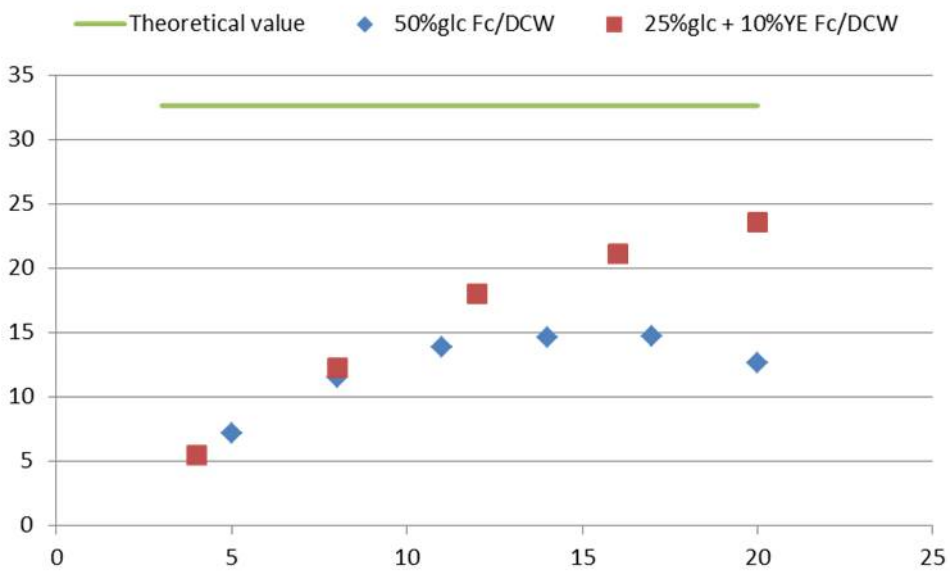


Figure 6: Ratio of Fc to DCW in two feeding cultures: 50% glucose and 25% glucose + 10% YE

Based on the expression data and biomass DCW during induction phase, it could be found that as time goes, before 15h, more and more resources are allocated to Fc production in both 50%glc and 25%glc+10%YE conditions, and compared to 50%glc, with 10%YE complement, more resources are allocated to Fc production (Figure 6). This inspires me to apply a new model to deal with the resources distribution between biomass and recombinant protein.

Scott M et al. proposed a new proteome resource allocation theory in their impressive Science paper in 2010. That is, considering a four-component proteome consisting of: 1) Φ_R : class R of mass fraction, containing all the ribosome proteins and their affiliates 2) Φ_Q : class Q of mass fraction, a fixed sector 3) Φ_P : class P of mass fraction, non-ribosomal-associated but growth-rate dependent (e.g., metabolic proteins) 4) Φ_U : unnecessary protein of mass fraction, e.g., recombinant protein

By definition:

$$\Phi_R + \Phi_Q + \Phi_P + \Phi_U = 1$$

And meanwhile based on empirical correlations found by the authors:

$$\Phi_P = \rho \times \frac{\lambda}{\kappa_n}$$

$$\Phi_R = \rho \times (r_0 + \frac{\lambda}{\kappa_t})$$

$$1 - \Phi_Q = \rho \times r_{max}$$

we could have

$$\frac{\lambda(\Phi_U)}{\lambda(\Phi_U = 0)} = 1 - \frac{\Phi_U}{\Phi_C}$$

where ρ is conversion factor which could be considered as constant, λ is specific growth rate, κ_n and κ_t are nutritional and translational capacity of the organism, and $\Phi_C = \rho \times (r_{max} - r_0) \approx 0.48$, $\lambda(\Phi_U)$ is the growth rate with Φ_U in the proteome, and $\lambda(\Phi_U = 0)$ is the growth rate without unnecessary protein in the proteome.

According to this theory, the growth reduction induced by heterologous protein expression is a simple consequence of proteome allocations. And more heterologous proteins, less growth rate. Our data of Fc and biomass corroborates this. In addition, considering mass fraction of the proteome in biomass is 55%-68%, then we could deduce theoretical maximal mass fraction of recombinant protein in biomass is 32.64% (Figure 6).

Rather than original recombinant constraints parameters to adjust the resources allocation between cell's objective to increase biomass and our goal to express protein, this new theory seems simpler and clearer. Here, before applying it to our bioprocess modelling, there still remains one issue to be addressed. In the beginning of induction, $\Phi_U = 0$, so $\lambda(\Phi_U) = \lambda(\Phi_U = 0)$, all resources would be allocated to the biomass growth, $\Phi_U \equiv 0$. So, we have to force the cell to invest proportional resources to the recombinant protein synthesis. This proportion would determine how fast the recombinant protein is synthesized in the initial time. In Fc case with 50%glc and 25%glc + 10%YE conditions, proportions are 0.2 and 0.3, respectively, which could result in a very good agreement.

2.6 Apply the model to defined and complex medium

```
library(deSolve)

## Set the constraints
lowbnd(model)[react_id(model) == 'EX_glc__D(e)'] = -9.58; ## for HNC47 with Fc
lowbnd(model)[react_id(model) == 'EX_o2(e)']=-11.5; # for HNC47 with Fc

## Add Fc synthesis reaction into the model
model <- addReact(model,id="RecombinantProtein",
           met=c('ala__L[c]','cys__L[c]','asp__L[c]','glu__L[c]','phe__L[c]','gly[c]',
           'his__L[c]','ile__L[c]','lys__L[c]','leu__L[c]','met__L[c]','asn__L[c]',
           'pro__L[c]','gln__L[c]','arg__L[c]','ser__L[c]','thr__L[c]','val__L[c]',
           'trp__L[c]','tyr__L[c]','atp[c]','adp[c]','pi[c]'),
            Scoef=c(-c(6,6,9,17,9,11,5,4,17,17,3,11,19,11,6,24,14,23,4,8),-4.306*224,964.544,965.544),
            lb=0,ub=1000,obj=0);

## Define FDM recipe, while we suppose NH4,Ca2+,cl-, fe2+, and cu2+ are abundant enough
substrateRxns = c('EX_glc__D(e)','EX_nh4(e)','EX_pi(e)','EX_so4(e)','EX_mg2(e)',
                  'EX_k(e)','EX_ca2(e)','EX_cl(e)','EX_fe2(e)','EX_mn2(e)',
                  'EX_cu2(e)','EX_zn2(e)','EX_mobd(e)'
                  );
initConcentrations = c(55.56,2000,67.6,63.08,2,67.6,1000,1000,2000,0.09171,1000,0.05912,0.01078);

## EX_ni2(e) and EX_cobalt2(e) both are required to support the growth of cell
lowbnd(model)[react_id(model) == 'EX_cobalt2(e)'] = -1000;
lowbnd(model)[react_id(model) == 'EX_fe3(e)'] = 0;
lowbnd(model)[react_id(model) == 'EX_tungs(e)'] = 0;
lowbnd(model)[react_id(model) == 'EX_ni2(e)'] = -1000;
lowbnd(model)[react_id(model) == 'EX_sel(e)'] = 0;
lowbnd(model)[react_id(model) == 'EX_slnt(e)'] = 0;

## 10% YE + 25% glucose for feeding during induction
## here, suppose YE is merely composed of amino acids
AA <-c('EX_ala__L(e)','EX_arg__L(e)','EX_asn__L(e)','EX_asp__L(e)','EX_cys__L(e)',
       'EX_gln__L(e)','EX_glu__L(e)','EX_gly(e)','EX_his__L(e)','EX_ile__L(e)',
       'EX_leu__L(e)','EX_lys__L(e)','EX_met__L(e)','EX_phe__L(e)','EX_pro__L(e)',
       'EX_ser__L(e)','EX_thr__L(e)','EX_trp__L(e)','EX_tyr__L(e)','EX_val__L(e)')
```

```
conc_aa_feeding <- 2.5*c(15.018,8.795,0.000,6.655,0.000,0.000,16.459,7.423,0.000,6.955,
+ 11.957,5.917,2.207,5.588,3.082,4.451,6.036,0.000,1.175,9.289)

## Define feeding component concentrations,
## where suppose O2 is always ample during the whole fermentation process
feedSubstrateRxns = c('EX_glc__D(e)','EX_nh4(e)','EX_so4(e)','EX_mg2(e)','EX_fe2(e)',
                      'EX_mn2(e)','EX_cu2(e)','EX_zn2(e)','EX_mobd(e)',AA);
feedConcentrations = c(2775.3,0.00924,40.23174,40,0.0719,0.09171,0.00881,0.05912,0.01078);
feedConcentrations = feedConcentrations/2;
feedConcentrations = c(feedConcentrations,conc_aa_feeding);

## This step runs dGSMEF, and will take a long time
#Ec_df2 <- dGSMEF(model,substrateRxns = substrateRxns,initConcentrations = initConcentrations,
#                initBiomass = 0.0142,u_fix = 0.15,x_ind = 31.63,
#                feedSubstrateRxns = feedSubstrateRxns,feedConcentrations = feedConcentrations,
#                yield_rate = 0.6586, # 25% glucose + 10% YE
#                ngam = 6.34,
#                initRatio = 0.70,
#                feedRate_ind = 2.8,
#                timeStep = 0.25,nSteps = 180,
#                rcd = TRUE,fld = TRUE);
```

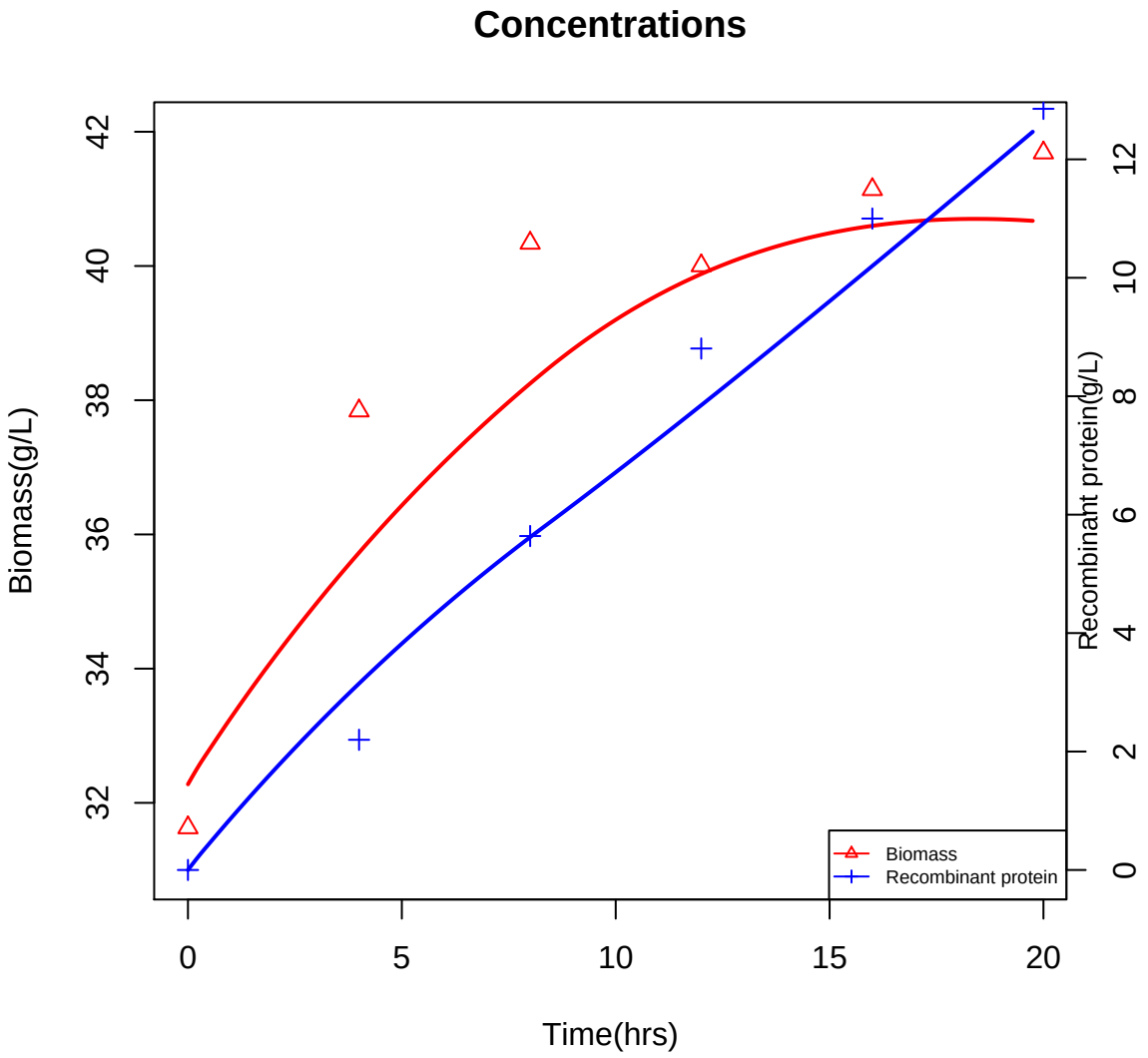
```
## compare predictions with experimental results during induction phase
locs <- Ec_df2@biomassVec >= 31.63
ind_time = Ec_df2@timeVec[locs][1]
plot(spline(Ec_df2@timeVec[locs]-ind_time,Ec_df2@biomassVec[locs], n = 201, method = "natural"),
     col = 2,main='Concentrations',xlab='Time(hrs)',ylab="Biomass(g/L)",ylim = c(31,42),type="l",lwd=2)

ind_time = 0
points(c(ind_time,ind_time+4,ind_time+8,ind_time+12,ind_time+16,ind_time+20),
       c(31.63,37.84164381,40.34110603,40.00462097,41.14031209,41.68691719),pch=2,col=2)

par(new=T);

ind_time = Ec_df2@timeVec[locs][1]
plot(spline(Ec_df2@timeVec[locs]-ind_time,
            Ec_df2@concentrationMatrix[Ec_df2@excRxnNames %in% "Protein"][locs]*25163.41/1000,
            n = 201, method = "natural"),axes=F,col = 4,xlab="",ylab="",type="l",lwd=2);

ind_time = 0
points(c(ind_time,ind_time+4,ind_time+8,ind_time+12,ind_time+16,ind_time+20),
       c(0,2.198356194,5.63889397,8.805379035,10.99968791,12.85308281),pch=3,col=4)
axis(4);
mtext("Recombinant protein(g/L)",side=4,cex=0.8);
legend("bottomright", c("Biomass","Recombinant protein"), col=c(2,4), lty=1, pch=c(2,3),
       text.font = 1.2, cex = 0.6);
```




```
## save flux distributions
Matrix <- matrix(Ec_df2@fluxMatrix,nrow=react_num(model),ncol=length(Ec_df2@timeVec),
                 byrow = FALSE);
subsystemsMatrix <- subSys(model);
subsystems_name <- colnames(subsystemsMatrix);
subsystems <- apply(subsystemsMatrix,1,function(x) paste(subsystems_name[x],collapse = ", "));

react_tmp <- printReaction(model,react=c(1:react_num(model)),printOut=FALSE);
react_equation <- unlist(strsplit(react_tmp,"\t"))[seq(2,2*react_num(model),2)];

Matrix <- cbind(react_id(model),react_name(model),react_equation,subsystems,Matrix);
colnames(Matrix) = c("ReactID\\Time(h)","ReactName","ReactEquation","Subsystem",
                     Ec_df2@timeVec);
write.table(Matrix,file="HNC47_flux_distributions.tab",sep="\t",row.names=FALSE);

## save concentration distributions
concMatrix <- matrix(Ec_df2@concentrationMatrix,nrow=length(Ec_df2@excRxnNames),
```

```
                    ncol=length(Ec_df2@timeVec),byrow=FALSE);
colnames(concMatrix) = col.names=Ec_df2@timeVec;
rownames(concMatrix) = Ec_df2@excRxnNames;
write.table(concMatrix,file="HNC47_conc_distributions.tab",sep="\t",col.names=NA);

## reduced costs
Matrix <- matrix(Ec_df2@redCostMatrix,nrow=react_num(model),ncol=length(Ec_df2@timeVec),
                 byrow = FALSE);
subsystemsMatrix <- subSys(model);
subsystems_name <- colnames(subsystemsMatrix);
subsystems <- apply(subsystemsMatrix,1,function(x) paste(subsystems_name[x],collapse = ", "));

react_tmp <- printReaction(model,react=c(1:react_num(model)),printOut=FALSE);
react_equation <- unlist(strsplit(react_tmp,"\t"))[seq(2,2*react_num(model),2)];

Matrix <- cbind(react_id(model),react_name(model),react_equation,subsystems,Matrix);
colnames(Matrix) = c("ReactID\\Time(h)","ReactName","ReactEquation","Subsystem",Ec_df2@timeVec);
write.table(Matrix,file="HNC47_redCost_distributions.tab",sep="\t",row.names=FALSE);
```

3. Metabolic flux analysis

3.1 PCA analysis

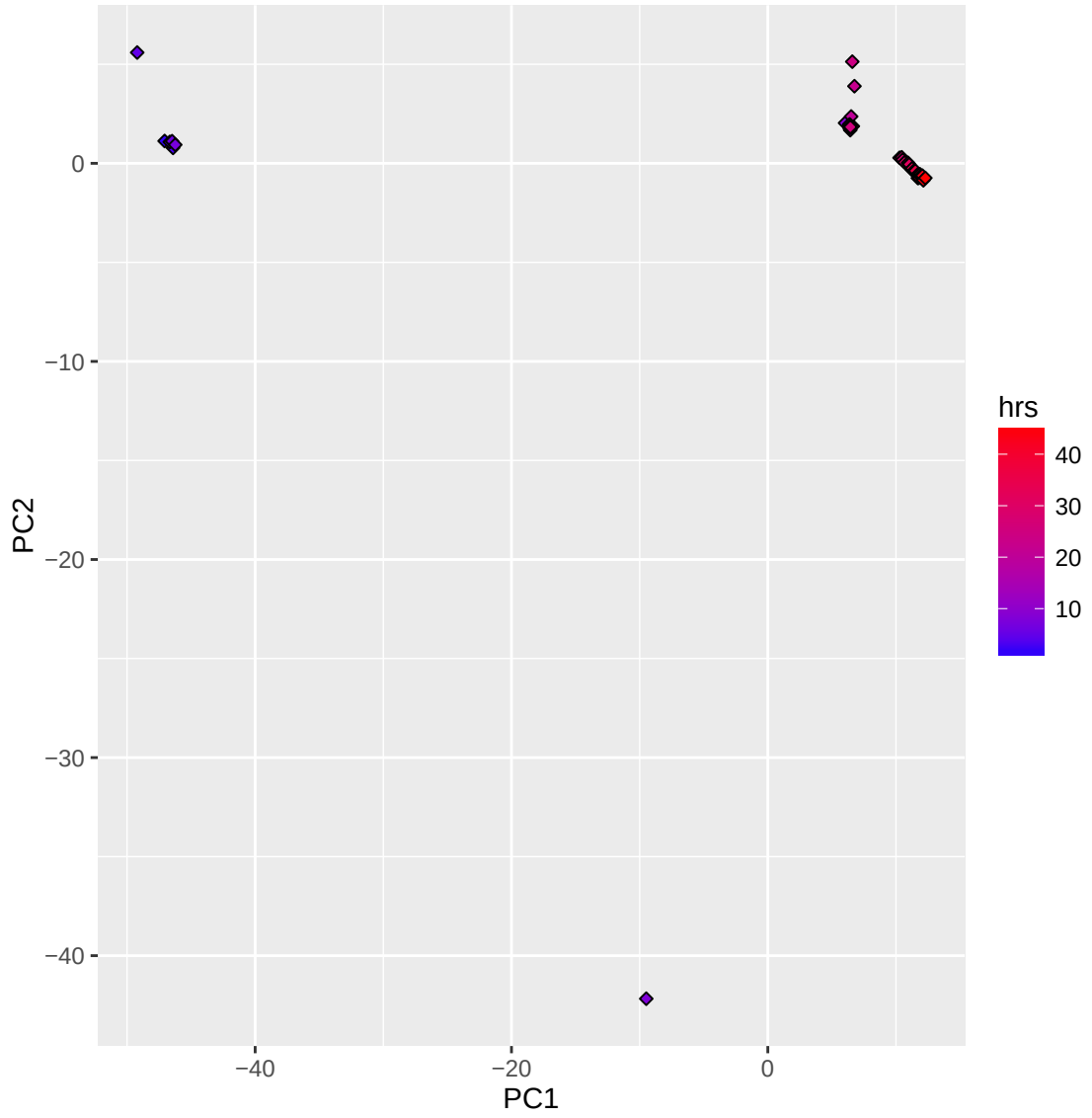
Analyzing the results of a genome-scale dynamic flux balance analysis presents a significant challenge since more than 2,000 variables (reaction fluxes) are available for each simulation. Principal Component Analysis (PCA) is a multivariate data analysis technique capable of making sense of such datasets. PCA is a projection method designed to display systemic variation and extract information from noise in a data matrix X by projecting points to a number of principal components. The information in many variables can be captured by a few principal components. Plots based on PCA show relationships between variables and between observations, indicating degree of correlation between variables. Before analysis, the input data is mean-centered and scaled to unit variance. Following r codes could generate such a plot visualizing differences between data points based on relative separation. One data point means a flux distribution at a specific time point during the whole fermentation process. There appears to be 3 major classes, just corresponding to 3 phases of fermentation. Actually, we could apply this PCA analysis to a more complicated case, e.g., metabolic variations over cell-cultures with different amino acid supplementation. From a bioprocessing standpoint, to some degree, we merely hope to increase production, without perturbing cell culture state. Because a significant change in cellular metabolism will in theory lead to potential changes in product quality attributes or undesirable side-consequences.

```
rawdata <- read.table("HNC47_flux_distributions.tab",header=TRUE,sep="\t");
## excluding exchange reactions
rawdata <- rawdata[-grep("EX_",as.character(rawdata[,1])),]
#rawdata <- rawdata[-grep("Biomass",as.character(rawdata[,1])),]
#rawdata <- rawdata[-grep("RecombinantProtein",as.character(rawdata[,1])),]
flux_names <- as.character(rawdata[,1])
flux_data <- rawdata[,6:ncol(rawdata)];

## to compare the flux distributions along with fermetation time
time_points = seq(4,ncol(rawdata)-5,4) # one point per one hour
flux_tmp = t(flux_data[,time_points])
flux_tmp = flux_tmp[,apply(flux_tmp,2,var) != 0]
# since more variables than samples, use prcomp instead
flux.pr <- prcomp(flux_tmp,scale=TRUE)
```

```
pr_matrix <- cbind(data.frame(flux.pr$x[,1:2]),time_points/4 )
colnames(pr_matrix) <- c("PC1","PC2","Time")

library(ggplot2)
p = ggplot(pr_matrix,aes(x=PC1,y=PC2)) + geom_point(shape=23,aes(fill=Time)) +
    scale_fill_continuous(low="blue",high="red")
# set legend text and title
p + labs(fill = "hrs")
```



3.2 Identify metabolic reactions that are shifted significantly

A typical idea to determine intracellular flux changes significantly is described as follows. First, reaction fluxes are sampled for two conditions. Subsequently, sample of flux differences is calculated by selecting random flux values from each condition to obtain a distribution of flux differences for each reaction. Finally, standardized reaction Z-scores are determined, which represent how far the sampled flux differences deviates from a zero flux change, or the background. Reaction scores can be used in visualizing perturbation

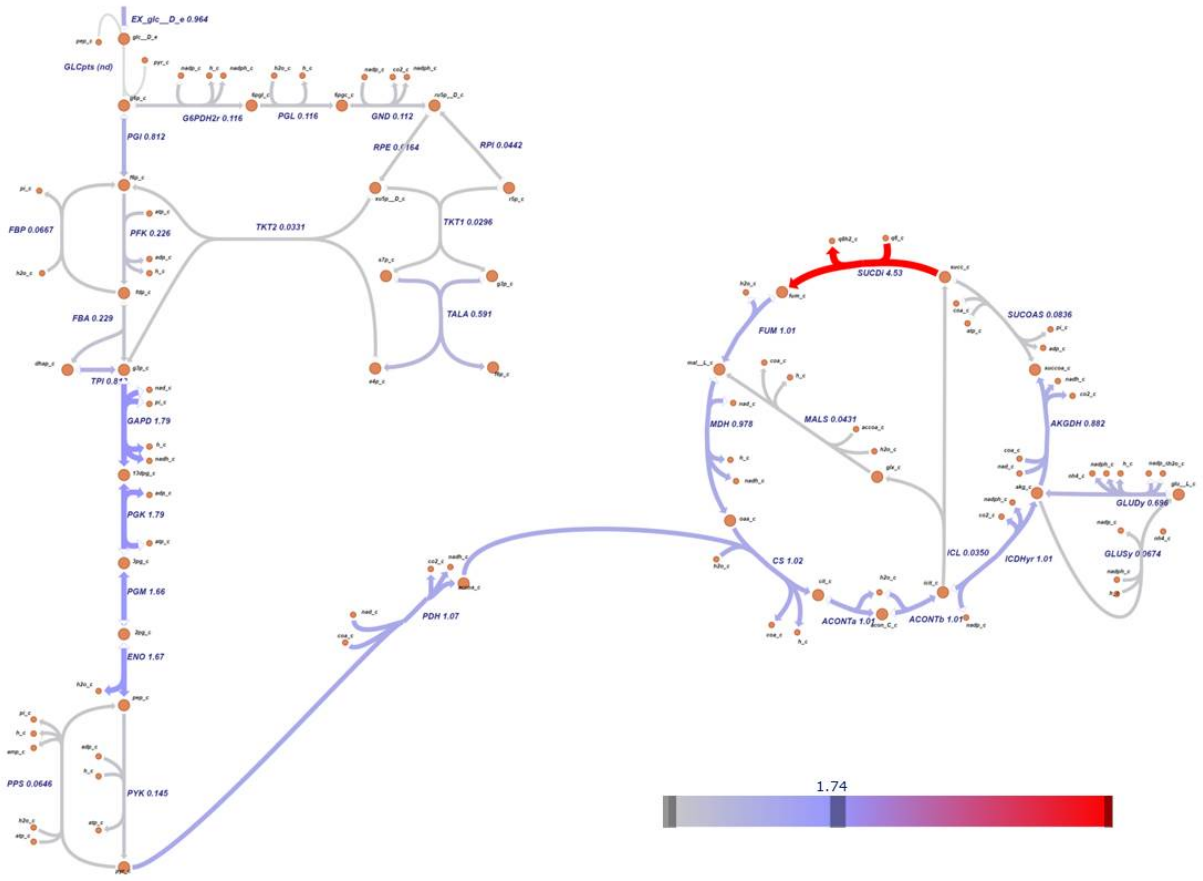


Figure 7: Genome-scale modeling revealed perturbations in glycolysis and TCA cycle activitity associated with changed feeding cultures after induction (phase 2). Escher is used for building, viewing, and visualizations of biological pathways

subnetworks and analyzing reporter metabolites and subsystems. For more details, please refer to the workflow (https://github.com/Kange2014/FluxShift/blob/master/time_course_fluxShift.ipynb)

This z-score-based analysis is carried out to determine the most significantly shifting fluxes between 50%glc and 25%glc + 10%YE feeding conditions. I select three different time points, 10h after batch (phase 1), and 5h, 10h after induction (phase 2, 3) for comparison. With YE replacement of 25%glc, majority of significantly shifting reactions are from glycolysis and TCA cycle, in particular for post-induction. Changes on other reactions in these two pathways although are not that significant, they also changed a lot, when compared to the background (Figure 7). Visualization of these perturbed reaction nodes brings about a striking commonality that many are NADH-producing reactions, such as GAPD, PDH, MDH, AKGDH, and those related to specific amino acid biosynthetic pathways (PGK to serine, ICDHyr to glutamate, MDH to aspartate).

3.3 PTM analysis

In *E. coli*, both methylation and acetylation can occur on recombinant proteins. For the former, S-adenosylmethionine (SAM) is reported to be the only methy-donor and thus, its concentration and related enzymes, e.g., methyltransferases, will affect the status of protein methylation. It is also reported that S-adenosylhomocysteine (SAH) is a competitive inhibitor of the methyltransferases, therefore the SAM:SAH ratio dictates the activity of the transferases. And because homocysteine could be converted from SAH and

to Met and then to SAM, circulating homocysteine levels correlate inversely with cellular SAM:SAH ratio, too. In addition, demethylation depends on both O2 and a-ketoglutarate. Among TCA cycle metabolites, a-ketoglutarate promotes demethylation, whereas other dicarboxylic acids, such as succinate and fumarate, are competitive inhibitors of the a-ketoglutarate-dependent demethylases.

For the latter, acetylation, acetyl-coenzyme A (acetyl-CoA) is the acetyl donor for all acetylation modifications. There are two kinds of acetylation modifications:

- Nα-acetylation: N-terminal acetyltransferases perform Nα-acetylation on the α-amino group of the first amino acid residue of nascent polypeptides. As a common co-translational modification, Nα-acetylation controls protein synthesis, stability and localization of the vast majority of eukaryotic proteins and is an irreversible reaction. RimI, Riml and RimJ of E. coli are reported to relate to this kind of modification.
- Nϵ-acetylation: both dynamic and reversible. By contrast, post-translational acetylation catalyses site-specific NÎµ-acetylation of the ϵ-amino group of lysine residues, which is a reversible reaction. The reversibility of Nϵ-acetylation makes this post-translational modification especially favourable in the regulation of proteins in response to metabolic changes. Two mechanisms have been identified: one is enzymatic, dependent on an acetyltransferase and acetyl-coenzyme A; the other is non-enzymatic and depends on the reactivity of acetyl phosphate. In *E. coli*, YfiQ (also known as Pka, Pla and PatZ) is the only known KAT (lysine acetyltransferase), whereas the NAD+-dependent sirtuin CobB appears to be the sole or predominant KDAC (lysine deacetylase).

Metabolic flux analysis could provide some interesting insights to these two kinds of PTMs in *E.coli*. By summing all producing reaction fluxes for these relevant metabolites, the sum flux for each metabolite could be estimated. Table 5 shows the flux ratio of SAM:SAH with 25%glc+10%YE conditions undergoes 12.4 at phase 1 to 16.6 at phase 2, and finally to 18 at phase 3, while with 50%glc the corresponding values are 11.9, 14.4 and 17.9, respectively. It seems that with 25%glc+10%YE, relatively more SAM and less SAH are synthesized. High SAM:SAH ratio apparently indicates high methylation rate. Meanwhile, low levels of homocysteine are also observed with 25%glc+10%YE. In addition, with YE, there may be another factor to further influence homocysteine levels. YE component vitamin B12 could induce a faster conversion from L-homocysteine to methionine by activating enzyme metH. Demethylation by a-ketoglutarate, in contrast, is less favored due to low turnover rates of a-ketoglutarate (Figure 8). Consequently, it sounds plausible to observe a relatively high methylation status with YE supplement because both methylation and demethylation environments incline to it.

Table 5. Cumulative producing fluxes for three key metabolites related to methylation in *E. coli*

metabolite	25%glc_10%YE p1	25%glc_10%YE p2	25%glc_10%YE p3	50%glc p1	50%glc p2	50%glc p3
ahcys_c (SAH)	4.30E-05	1.39E-05	1.03E-05	5.62E-05	2.27E-05	1.27E-05
amet_c (SAM)	0.000532983	0.000231416	0.00018513	0.000668798	0.00032725	0.000227168
SAM:SAH	12.4	16.6	18	11.9	14.4	17.9
hcys___L_c (homocysteine)	0.010198546	0.003803588	0.002729702	0.019674748	0.0099403	0.006688996

Above metabolic flux analysis results have already shown that many significant shifting reactions are NADH-producing reactions. As a result, it is interesting to examine the cumulative producing flux of NADH, especially when we consider it also relates to protein acetylation levels. From Figure 8, it could be observed that both acetyl-CoA and NADH are produced less with 25%glc+10%YE condition. Low levels of acetyl-CoA indicates low level of acetylation, and low NADH means higher NAD+/NADH ratio. This NAD+/NADH ratio changes seem likely to regulate the protein acylation state, with deacylation occurring when the NAD+/NADH ratio is high and acetylation being favored when the NAD+/NADH ratio is low. Low acetylation and high deacylation with 25%glc+10%YE could explain why under this culture condition, less acetylation occurs on the recombinant protein Fc. In addition, non-enzymatic dependent acetyl phosphate doesn't show significant difference between these two conditions (data not shown). However, it is still unclear which type of acetylation

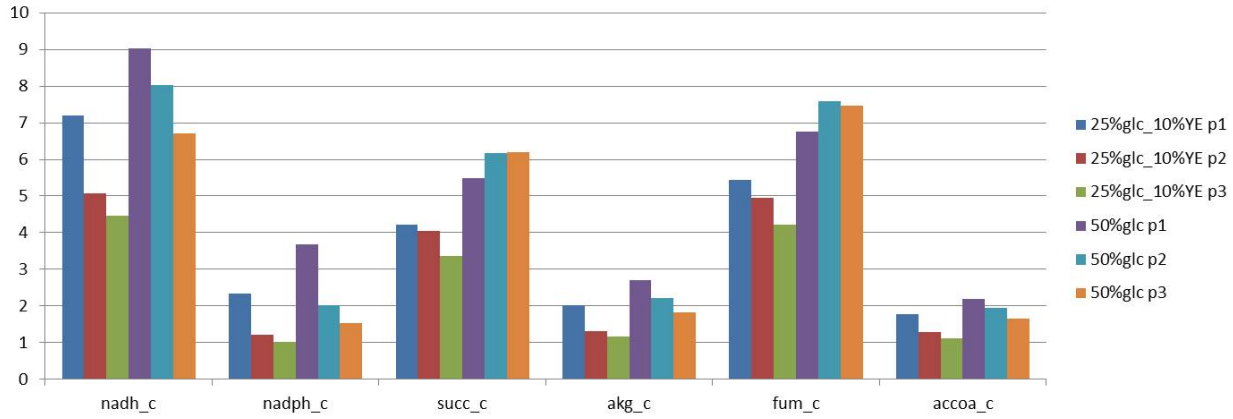


Figure 8: Cumulative producing fluxes for key metabolites related to methylation and acetylation in E. coli

(Nα-acetylation or Nϵ-acetylation) is preferred. More efforts are required to investigate it.

4. Reference

1. O'Brien EJ, et al. Cell 2015, 161 (5), 971-987;
2. King ZA, et al. Nucleic Acids Res 2015, 44 (D1);
3. Pramanik J, et al. Biotechnol. Bioeng. 1997, 56 (4), 398-421;
4. Meadows AL, et al. Metabolic Engineering 2010, 12 (2), 150-160;
5. Mahadevan R, et al. Biophysical Journal 2002, 83 (3), 1331-1340;
6. Scott M, et al. Science 2010, 330(6007):1099-102.