Chapter 11

Benefit and cost of a protein

Manish Kushwaha, Wolfram Liebermeister, Elad Noor, Kirill Sechkar, Diana Széliová

Chapter overview

- How are global resource allocation and cell fitness reflected in local choices such as determining the expression level of a single protein? Separating between the cost and benefit of a specific protein's expression helps describe this local optimization quantitatively.
- Effective costs and benefits of a single protein may be defined operationally, as measured increases or decreases in cell growth. For enzymes, the factors that determine cost and benefit include enzyme efficiency and protein cost (size, energy, amino acids, ribosomes needed to make proteins, as well as all kinds of opportunity costs).
- Marginal cost (or benefit) is defined as the derivative of the cost (or benefit) curve, i.e. the effect of small changes in expression. At the point of maximal fitness, the marginal cost and marginal benefit must be equal.
- We describe a simple cost-benefit model where benefit is defined as the production flux and cost as the enzyme level. Based on the same idea, we then present empirical definitions and measurements of the costs and benefits of the lac-operon in *Escherichia coli*.
- We discuss predictions from constraint-based and mechanistic cell models and compare them to the empirical definitions.
- In an appendix, we describe methods for measuring transcription, translation, and growth rates, as well as for controlling expression to observe costs/benefits for different protein levels.

11.1 Effects of protein expression on cell growth

11.1.1 How do changes in a protein level affect cell growth?

Protein production in cells is a complex process that involves transcription, translation, modifications, and transport, and it depends on various other processes. When protein expression levels are changing, this will affect ribosome demand and metabolic production fluxes. Proteins also have forward effects, not only by performing various cellular functions but also by competing for space among each other and with other compounds. Changes in protein levels may have a variety of indirect effects, including rearrangements of gene expression, which depend on regulatory mechanisms and will impact the cell-wide allocation of resources. In models, these changes may either be described physically–as a result of regulatory mechanisms–or based on optimality principles. Our models from previous chapters are aimed at predicting such rearrangements.

Having learned how to model resource allocation in cells, we now ask a simple question: When a cell expresses a heterologous protein or overexpresses (or underexpresses) a native protein, how will this change its growth rate?

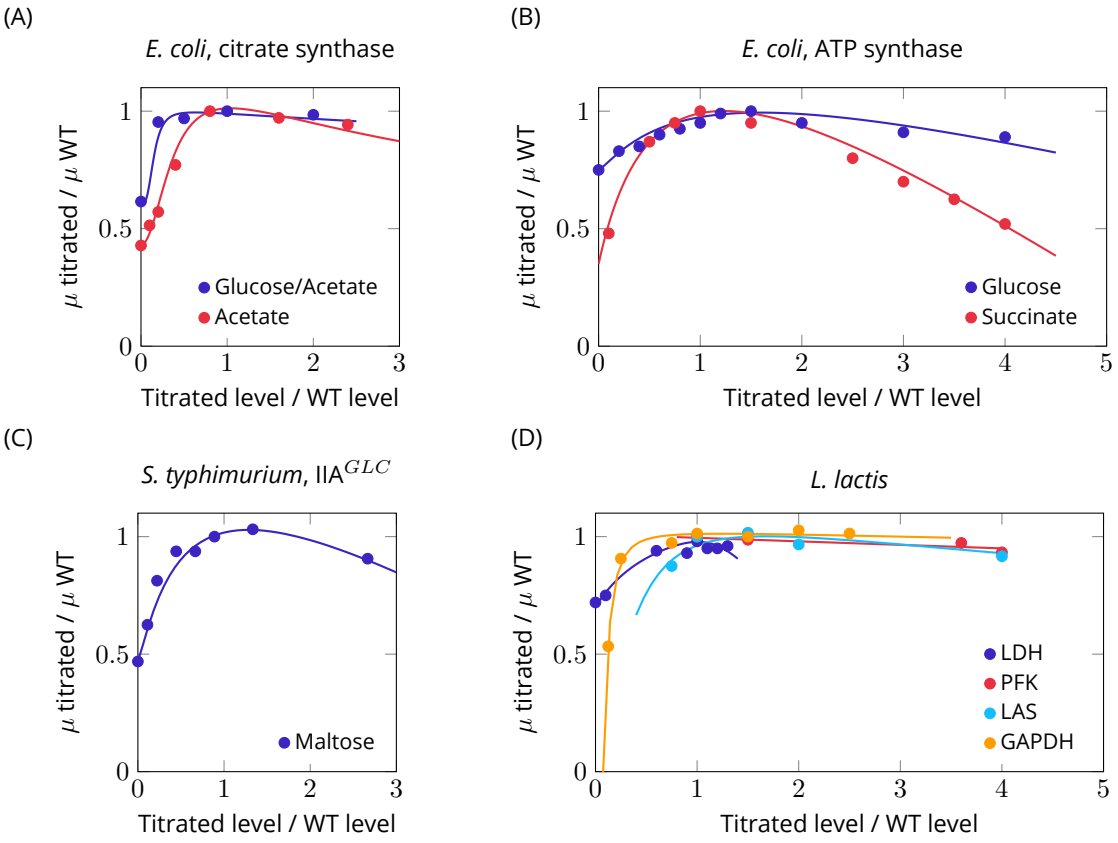


Figure 11.1: Growth rate effects of protein expression in different bacterial species – The panels show different examples of optimal enzyme expression. (A) Citrate synthase in *Escherichia coli*. (B) ATP synthase in *E. coli* (C) PTS transporter system (glucose-specific subunit IIA) in *Salmonella typhimurium*. (D) Several glycolytic enzymes in *Lactococcus lactis*. Figure redrawn from [1] with kind permission by the authors. The ATP synthase data stem from [2]. In all cases, "Titrated levels" refer to proteins under the control of an IPTG-inducible promoter.

Since cell growth depends on many cellular processes, the answer to this question can be quite complex. However, according to simple economic logic, each protein under each given condition should exhibit an optimal expression level. If a protein level is too low, the protein cannot sufficiently exert its function, and if it is too high, the protein consumes too much of the cellular resources, which will then be lacking for processes elsewhere. As often in life, the optimum is a compromise, a "sweet spot" at which cells reach a maximal growth rate (or optimize some other relevant fitness objective). At this optimal protein level, any increase or decrease would result in a decrease in the growth rate. Of course, we may wonder: will cells actually achieve this optimal protein level?

We can test this in experiments. Figure 11.1 shows the expression of a number of proteins in different bacterial species. In this set of experiments, the amount of each selected protein was varied by placing its coding gene under the control of an inducible promoter, and cell growth rates were measured for several expression levels. All measured protein/growth curves were negatively curved (i.e. concave), where the maximum growth rate was found at an intermediate protein expression level. Moreover, the wild-type level of each protein–that is, the level in natural, non-modified cells–was close to its optimal level. By setting the enzyme level to this optimal value, wild-type cells often come close to the maximal possible growth rate. Figure 11.2 shows this for one specific protein, ATP synthase in *E. coli*, under a large range of growth conditions. The close match between the observed and maximal growth rates suggests that the regulatory mechanisms behind this protein have evolved to adjust the protein levels to support maximal growth. And this protein is not an exception. Wild-type levels of other proteins have been found to be growth-optimal as well, including those of some efflux pumps [3] and the enzyme methionine synthase [4].

Of course, this principle–wild-type protein levels are the protein levels that maximize growth–does not hold for all proteins and all types of cells. In yeast cells, for example, some protein levels were found to be growth-optimal while

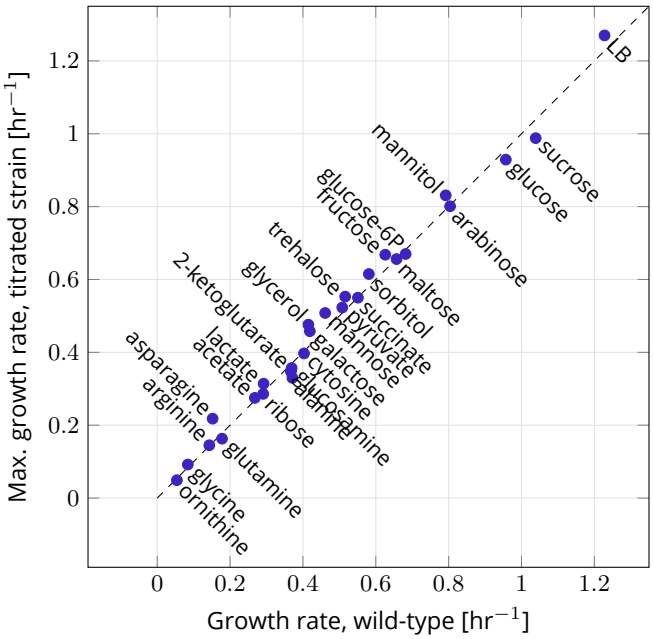


Figure 11.2: Wild-type abundances of ATP synthase in *E. coli* lead to near-optimal growth rates – The plot compares the growth rate of wild-type cells in different environments to the corresponding maximal growth rate achievable by varying the abundance of ATP synthase. Only in four of the 27 growth conditions shown, the wild-type growth rate of *E. coli* deviates by more than 10% from the maximal growth rate. Figure redrawn from [1] with kind permission by the authors. The ATP synthase data were taken from [2].

others were not, and this also depends on the growth conditions [5, 6]. But given the complex functioning of real cells, it still remains interesting to see what would be the optimal protein levels in theory, and by what general logic we can understand them.

The growth rate effects of proteins are related to various relevant questions:

1. **Understanding protein usage and protein levels in naturally evolved cells**. What is the best expression level for each protein in wild-type cells (e.g. the titration levels that maximize the functions in Figure 11.1)? And what are the conditions under which each single protein should be expressed?
2. **Understanding selection pressures on protein levels, which shape evolution**. How do deviations from the optimal protein levels affect the fitness objective? For example, in Figure 11.1, how strongly do the curves decrease when deviating from the optimal point? Knowing the curvature, we can get an idea about the cost of fluctuations and, therefore, flexibility in protein expression.
3. **Expressing heterologous pathways in cells in biotechnology**. Optimal choices of protein levels are also important for biotechnology. Optimizing protein levels for maximal production of valuable compounds or for maximal growth rates (e.g., by expression via controllable gene promoters) is a relevant biotechnological task and an example of cellular economics. Whenever proteins are expressed artificially, it is important to understand – and possibly limit – the effects on cell growth. A large growth deficit will slow down the reproduction of the cells and, therefore, total production. Moreover, a lower growth rate may lead to evolutionary selection against the engineered strain: mutants with disabled heterologous genes may experience lower growth defects and therefore take over the population by outcompeting the engineered cells.

But the question remains: How do these protein/growth curves arise? What determines their shape, and how do they differ between proteins? In this chapter, we consider the growth rate effects of proteins in detail.

Following the economic perspective of this book, our main focus is not on how protein levels are regulated in cells but on how they contribute to the cell's functioning. That is, we do not inquire about the mechanistic causes but about the *incentives* for increasing or decreasing a protein level. There are several reasons for looking at cells in this way. First, if there is an evolutionary selection for high growth rates, the resulting regulatory systems will support

optimal ”expression programs”. Second, even if we don’t fully understand how regulation systems function, we may use optimality principles as a way out, hoping that they can serve as a good approximation.

11.1.2 Protein cost and benefit

The curves in Figure 11.1 may look simple, but they may reflect complex rearrangements in the entire cell. Protein production is a complex process that involves transcription, translation, modifications, and transport and depends on various other processes. When protein levels change, all these processes may be adapted, the demand for ribosomes changes, and metabolic production fluxes may be rerouted. Proteins typically serve a specific function, and changing their abundance perturbs this function, too. Moreover, there are various indirect effects; for example, due to the competition for cellular space, which is crowded with other proteins, membranes, polymers, and small molecules (see Section 2.6.1). Mechanistically, these cell-wide rearrangements depend on regulatory mechanisms and impact the allocation of resources. Accordingly, there are two different ways to describe this reallocation in models: either physically – that is, as resulting from regulation mechanisms – or based on optimality principles. Our models from previous chapters can be used to predict such rearrangements.

One way to describe the effects of protein levels on cell growth is to assume that each protein has a benefit and a cost. On the one hand, we assume that a protein contributes to the functioning of the cell and that its presence has a positive effect on growth. On the other hand, protein production and maintenance require resources, which puts a burden on cell growth. These resources include precursors, energy, and labor time of the biosynthesis machinery. Furthermore, packing the cellular volume with more proteins would decrease the diffusion rate of small molecules in the cell (slowing down metabolism as a whole), which means that the space proteins occupy is yet another limited resource. The more abundant a protein is in a cell, the less space and material remains for other molecules that also perform important functions.

Below, we assume that any protein in a cell has a cost and that protein expression must be justified by a benefit, or the protein will not be expressed. At the optimal expression level, cost and benefit must be “balanced” such that any small changes in protein levels are fitness-neutral; that is, their additional costs and benefits cancel out. In principle, such costs and benefits can be studied by controlling the abundance of a protein via an external regulator, for example a compound that can activate the protein’s gene promoter, and measuring the resulting cell growth rate (Fig. 11.3(A). An “idle” protein, which provides no benefit to the cell in the current conditions, will only contribute a cost. The higher its expression level, the more slowly the cell will grow, and the optimum growth rate is reached when the protein is not expressed (Fig. 11.3(B)). In the case of a “useful” protein, there is also a benefit; when the expression level becomes very small, the benefit will show a sharp drop, with leads to an optimum at some intermediate expression level (Fig. 11.3(A)). Following this idea, and using such experimental data, we may split a protein’s growth effects into separate cost and benefit terms, hoping to understand each of these terms through simple considerations about the enzyme’s physical properties such as its molecular mass, size, or enzyme efficiency.

Notions of protein costs and benefits can be defined in different ways, depending on their purpose:

1. **Empirical definition and measurements**. A common operational definition is based on experiments in which the expression level of a protein is externally controlled, and the resulting changes in cell growth rate are measured. By comparing experiments in which beneficial and costly effects are varied, one can try to disentangle them to better understand what causes them in the cell.
2. **Model simulations**. Instead of real experiments, we may also perform computer experiments based on models. The models described in the previous chapters 7 and 10 allow us to simulate the change in enzyme levels and compute the resulting achievable growth rate.
3. **Theoretical definition and calculation**. Aside from numerical calculations, mathematical models can help us understand general principles (see box “Qualities of a model” in 5). If we made a model and know what assumptions went into it, we can understand it more easily than a real cell. Some model properties can be analyzed without simulations: metabolic control analysis, for example, can reveal general principles of optimal enzyme allocation and how it should be adapted upon an external change.

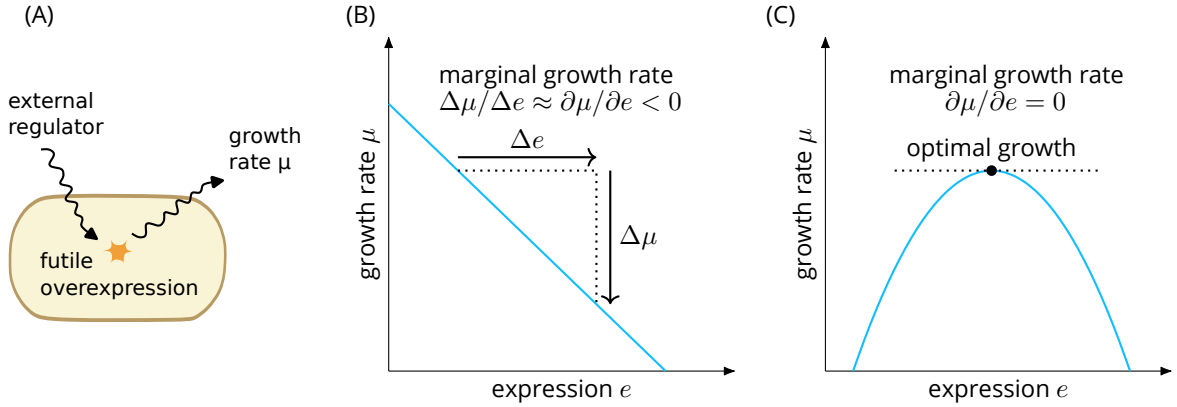


Figure 11.3: Quantification of protein/growth rate effects – (A) An externally controlled expression of a protein (with expression level e) influences the cell growth rate μ. (B) Forced expression of an idle protein decreases the growth rate. In wild-type cells under a selection for growth, such a protein should not be expressed. At small expression levels, the growth rate decreases approximately linearly (first-order effect). (C) In the case of a beneficial protein, the growth rate becomes maximal at some intermediate expression level, and small variations around this level make the growth rate decrease. Near the optimal point, the curve can be approximated by a quadratic curve (second-order effect). In the optimal point the curve has a zero slope, and small increases or decreases of the protein level would leave the growth rate almost unchanged (compare Figure 11.1).

In all three cases, the aim is to split cell-wide fitness effects into simple cost and benefit contributions attributed to a single protein. To justify this splitting, we often need to assume that protein changes are small compared to the entire proteome. This allows us, in many cases, to model protein costs and benefits using linear or quadratic approximations.

In Section 11.2 of this chapter, we consider cell growth as an example objective and describe how the influence of protein expression can be measured. You can find more details about experimental methods in the appendix. In Section 11.3, we introduce a simple cost/benefit model for a single enzyme; we then apply it to the Lac operon in *E. coli*, for which protein costs and benefits have been measured (Section 11.4). Finally, we discuss some general principles relating empirical costs and benefits to (direct or indirect) mechanistic causes, show how protein/growth curves can be predicted using the models discussed earlier in the book (Section 11.5), and ask how protein expression is regulated mechanistically and how this may lead to growth-optimal states (Section 11.6).

11.2 Protein levels and growth rate: mechanisms and observations

11.2.1 Protein production and the necessary resources

To understand how protein levels affect cell fitness, we first need to know the steps in protein production and the necessary resources. The amino acid sequence information for making a protein is (usually) encoded in DNA. DNA is transcribed to messenger RNA (mRNA) and then translated to protein. These processes require three basic ingredients: precursors, molecular machines that assemble these precursors, and energy to power these processes. The precursors for transcription are nucleotides, and for translation amino acids. These can be synthesized by the cell itself or be taken up from the medium. Transcription and translation are catalyzed by large molecular machines, such as RNA polymerases and ribosomes. Finally, these machines require energy in the form of ATP or GDP that the cell must produce (Figure 11.4).

Overall, protein translation is the most costly process. It consumes about four ATP molecules per amino acid, whereas transcription uses about two ATP molecules per nucleotide (BioNumbers ID 105375). Additionally, the catalysts involved in translation are larger; in bacteria, ribosomes are 2700 kDa (BioNumbers ID 100118), compared to polymerases with 480 kDa (Bionumbers ID 104925). Furthermore, an mRNA molecule is used to make multiple proteins.

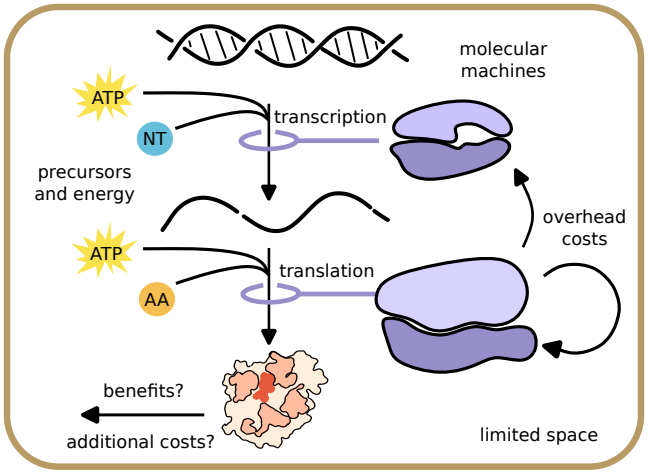


Figure 11.4: Steps in making a protein (see also Box 2.A in Chapter 2) – 1. Transcription: DNA is transcribed to mRNA by RNA polymerase, which requires nucleotides (NT) and energy (typically ATP or GTP). 2. Translation: mRNA is translated into a protein by the ribosome, using amino acids (AA) and energy. These processes are constrained by the available space and require additional indirect costs, as the molecular machines catalyzing these processes need to be synthesized by the same processes, necessitating even more machines. After the protein is made, it can be beneficial for the cell or cause additional costs (e.g. costs of folding and degradation of misfolded proteins, membrane stress, by-product accumulation), depending on the type of protein and environmental conditions. For more background about biological machines and enzymes, see Chapter 2.

Due to these cost differences, the translation machinery occupies a larger portion of the cell's proteome than transcription machinery (roughly 5-10 times larger in *E. coli*, see proteomaps.net [7]).

All of these processes contribute to the overall cost of protein production, with synthesis costs primarily determined by protein size. However, synthesis alone does not account for the full energetic burden. Proteins often require additional steps such as chaperone-assisted folding, post-translational modifications, or transport to specific cellular compartments, each of which consumes extra energy. In some cases, proteins can also impose toxic effects that further increase cellular energy demands. For instance, microbial production of biofuels can disrupt membrane fluidity and permeability, leading to the leakage of essential molecules such as ATP and ions [8].

Additionally, we need to consider that to make more protein, a cell also needs more molecular machines, which, in turn, requires even more molecular machines to synthesize themselves. However, the space in the cell is limited. In fact, as we have seen in Chapter 2, the protein content per cell is quite constant across conditions, but the relative composition changes. That means that to make more of one protein, the cell needs to reduce the amount of other (potentially beneficial) proteins – which can reduce growth. The limited space does not only apply to the cell as a whole but also to cellular substructures. For example, the abundance of membrane proteins is limited by the membrane area, and the abundance of mitochondrial enzymes is limited by the mitochondrial volume.

11.2.2 Cell-wide effects of protein expression: costs and benefits via different routes

To understand how a single protein can affect cell growth, we need to trace its effects on other processes in the cell. A change in a single protein level may imply – that is, either cause or require – rearrangements in the entire proteome. If the total protein amount in a cell were constant, then increasing the level of one protein would cause the levels of other proteins to decrease, with secondary effects elsewhere in the cell. In contrast, if the total protein amount can vary, an increased protein level may lead to an increase in the total protein amount, which entails adjustments in the protein production machinery (polymerases, ribosomes, available charged tRNA). This, in turn, has cell-wide effects. In both cases, if a change in protein levels affects the growth rate, this will entail various other adjustments (e.g. changes in ribosome numbers, with all kinds of secondary effects). To describe the cost of a protein, including all these effects

Experimental methods 11.A Challenges in quantifying protein expression

To quantify the fitness effects of a protein in real cells, one may enforce changes in protein levels and plot these levels against the resulting changes in growth rate. However, in experiments we cannot control protein concentrations directly. To vary it indirectly, we can modulate gene copy number, transcription, or translation, for example, by titrating an inducer (e.g., IPTG with *lac* promoter) or using promoters with different strengths. The inducer concentration or promoter activity is then used as the input variable for cost and benefit calculations [2]. However, these measurable variables may not correspond exactly to protein concentrations because multiple processes contribute to the final protein levels. These processes can become saturated or affected by regulation [9], and these effects are gene-specific [10]. There are also cases in which gene expression correlates with inducer concentration at the population level, but individual cells show "all-or-none" expression. This means that instead of each cell gradually increasing expression, only the proportion of fully induced cells rises [11], so the cell models used in this chapter would not even apply.

So how can we make sure that our calculations are done with accurate protein amounts? One solution is to measure the absolute level of our protein, for example with quantitative mass spectrometry. However, these methods are technically challenging – the measurement error are large and vary between proteins [9]. An alternative is to measure relative protein expression, for example using gel electrophoresis as done by Dekel and Alon [12] (see Section 11.4).

as "baggage costs" or opportunity costs of the protein, we usually assume that these costs are *comparable between proteins* and we can approximate them by empirical or heuristic protein cost functions.

To see how protein expression affects cell growth, we need to consider the different ways in which a protein – by its presence or changes in abundance – influences the rest of the cell. If we consider a cell as a whole and zoom in on a single protein species, we notice a number of direct connections between the protein and other cell processes. These "direct routes of action" concern (1) How the protein is made, including the need for resources such as amino acids, energy carried by ATP, molecular machines (including ribosomes and chaperones), and mRNA templates. In eukaryotes, transporting proteins to the right place in the cell may require transporters and larger infrastructures, such as the endoplasmic reticulum. (2) The space that the protein occupies (thus reducing the available space for other cell components). (3) Running costs, for example, ATP consumption in the case of ion transporters or flagella proteins. (4) The protein's direct function, for instance, is the catalysis of a metabolic reaction.

It is through these connections that our protein influences other parts of the cell. One way to predict how these influences play out when it comes to cell growth is to use detailed whole-cell models. But how can we understand these effects intuitively, and how can we relate them to cost and benefit terms? A changing protein level will change all the direct demands (for example, ribosome demand and space demand). To describe these demands, we now make a simplifying assumption. While different proteins require resources in different amounts, e.g. due to their different molecule sizes, the way in which these resource demands affect the cell (e.g. by higher ensuing ribosome demands) do not depend on the type of protein. If this is true, then knowing the cost factors of one protein will also tell us about the cost factors of other proteins (see Box **??**). Hence, the costs of proteins are assumed to depend only on general factors such as protein size, composition, lifetime, or localization. On the benefit side, unlike protein costs, protein functions differ largely between different proteins (for example, among the thousands of proteins in a cell, only one may be able to catalyze some specific reaction) – with important consequences for the entire cell!

In practice, it is common to summarize all cost factors of a protein into one cost term associated with the required resources and proportional to the protein's abundance, and to describe the protein's function by one protein-specific benefit term.

11.3 Cost-benefit balance for an enzyme level

If expressing a protein causes costs and benefits in the cell, then how can we find the "sweet spot", the expression level where fitness is maximized? And how will this optimal value change when external conditions are changing?

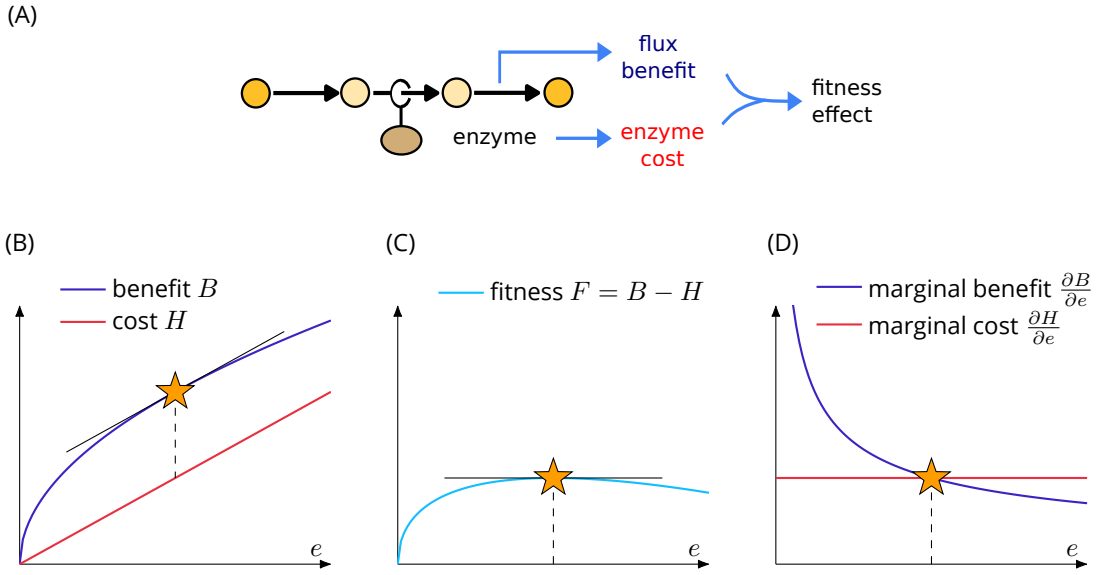


Figure 11.5: Optimal enzyme level in a metabolic reaction – (A) The enzyme influences the pathway flux via its catalyzing reaction. All other enzyme levels are fixed. In an optimality problem, we define cell fitness as a difference between a benefit proportional to the pathway flux and a cost proportional to the enzyme level. Internal metabolites are shown in yellow, external metabolites are shown in orange. (B) Cost (red) and benefit (blue) as functions of the enzyme level. The point of maximal fitness (benefit minus cost) is marked by an orange star. (C) The fitness, as a function of the enzyme level, has an optimum point in which the curve has a vanishing slope. (D) The derivatives of cost and benefit functions, as functions of the enzyme level, are called marginal cost and benefit. In the optimal point, marginal cost and benefit must be equal. We can already see this in (B), where the slopes are equal in the optimal point.

When external conditions are changing and make a protein more beneficial, it stands to reason that the cell should have more of it. But why exactly? And how are greater advantage and greater quantity related? A way to think about this is "marginal economics", that is, considering the costs and benefits of small changes around a given state of interest. The optimal point, where any change decreases the fitness, is also a point in which *very small* changes have no fitness effect: since the slope of the curve is zero, moving a bit to the left or right will lead to almost no change in the growth rate. We can also see this from the protein/growth curves in Figure 11.1. How is this related to costs and benefits? If the fitness function is benefit-cost difference, then its slope (called "marginal fitness") is the difference between the slopes of the benefit and cost curves, called marginal benefit and cost. In the optimal point, where the slope of the fitness function must be zero, marginal cost and benefit cancel out: this is what we mean by saying that "cost and benefit are in balance" (see Box **??**). In this logic, the "sweet spot" is a point where an infinitesimal change would have equal cost and and benefit effects, so it this point, there is no incentive to increase or decrease the enzyme level.

11.3.1 Cost-benefit balance of a metabolic enzyme – a simple additive model

To see a concrete example, let us consider a metabolic pathway containing an adjustable enzyme as shown in Fig. 11.5 (A) and assume that the enzyme level is optimized [13]. We further assume that the pathway contributes to cell fitness in two ways, described by separate fitness terms. The benefit term B depends on the pathway flux J, while the cost term depends on the total pathway enzyme e_{tot}, and the fitness is given by the difference $F = B - H = b_J\, J - \gamma\, e_{\mathrm{tot}}$ with constant prefactors b_J and γ. The cost H is proportional to the enzyme level and may represent opportunity costs of the enzyme due to its ribosome and space demand. Note that fitness is measured here in arbitrary units, while the units of the coefficients b_J and γ must such that cost and benefit functions have the same unit. For example, if the cost function denotes enzyme mass concentrations (with γ, in this case, being the enzyme molecular mass), then b_J must represent the benefit measured as an enzyme mass concentration per flux.

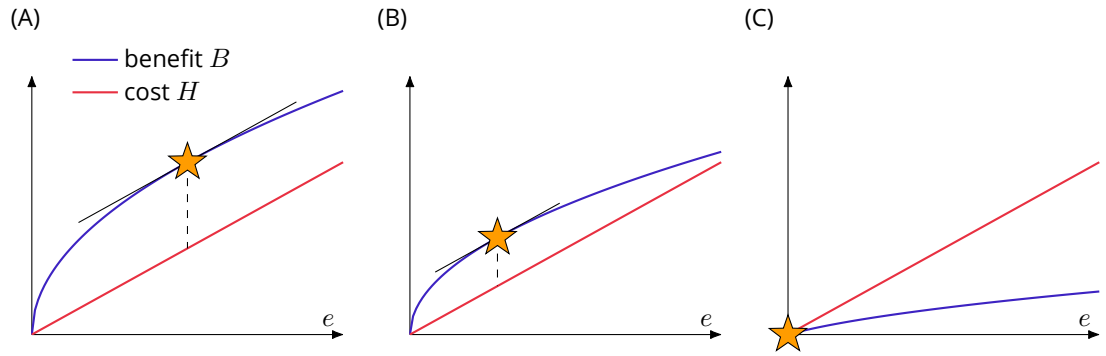


Figure 11.6: Changing the balance between cost and benefit shifts the optimal protein expression level – In the cost-benefit model from Figure 11.5 (left panel), we decrease the benefit weight, leading to lower optimal expression levels (center) and eventually to a zero expression (right).

How can we find the optimal enzyme level, and how will this optimum change as external conditions are changing? We treat our enzyme level e as a choice variable, keep all other enzyme levels constant and ignore their constant cost, and write the fitness as a function of our enzyme level alone:

$$F(e) = \underbrace{b_J\, J(e)}_{B(e)} - \underbrace{\gamma\, e}_{H(e)}. \tag{11.1}$$

In a simple kinetic model, the flux $J(e)$ will vary between 0 (at $e = 0$) and some maximal value v_∞ (at $e \to \infty$), which depends on the other enzyme levels. To obtain mathematical conditions for an optimal enzyme level e, we first consider a case in which the optimal enzyme level is larger than 0. In this case, the derivative $\mathrm{d}F/\mathrm{d}e$ must vanish in the optimum point, so our protein must have a marginal fitness of zero. This means that the slopes of the cost curve $B(e)$ and benefit curves $H(e)$, called marginal cost and benefit, must balance out:

$$\underbrace{\frac{\partial B}{\partial e}}_{b_J \frac{\partial J}{\partial e}} = \underbrace{\frac{\partial H}{\partial e}}_{\gamma}, \tag{11.2}$$

This optimality criterion holds for any curve shapes, for other types of choice variables aside from protein levels, and also in other types of model (e.g. where benefits emerge from a complex cell model, with an unknown, but mathematically defined dependence on some choice variables).

How will an optimal enzyme level change as conditions are changing? For example, how should it change if an enzyme remains equally costly, but becomes less beneficial? In our model, let us assume that the benefit function is scaled down by a tunable factor while the cost function remains unchanged. If the original optimal enzyme level e^* was positive, then rescaling our benefit function changes its relative weight compared to the cost term: the shape of the fitness function changes and the optimum enzyme level shifts to the left (see Figure 11.6). The more we decrease the benefit, the lower the optimum level will be, until the optimum level hits 0. If we further decrease the benefit weight, the slope of the fitness function at zero expression becomes negative, and expressing the enzyme would not pay off anymore. Since enzyme levels cannot be negative, we obtain a boundary optimum at $e = 0$. When we give the benefit function a weight of 0 (assuming that the pathway flux does not contribute to the cell's benefit at all), we obtain an "idle" protein, which has a cost, but no benefit. According to our optimality model, such proteins should not be expressed. An example of an idle protein is the Green Fluorescent Protein (GFP), which is sometimes cloned into cells for measurement purposes and puts a burden on the cell [14].

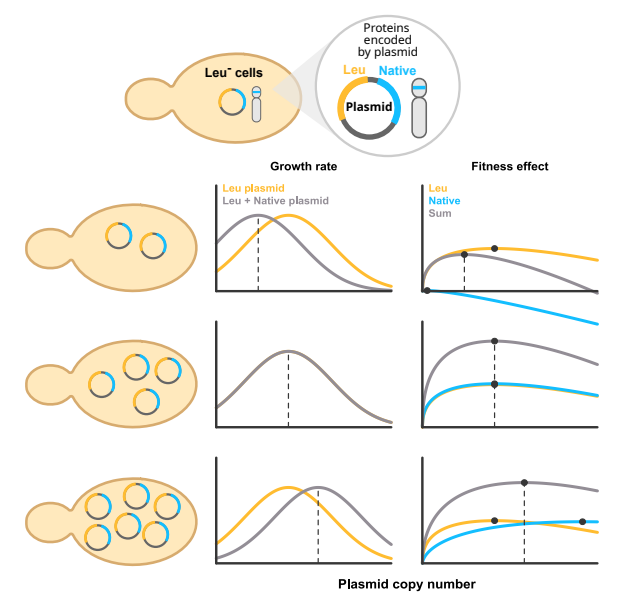
Even if our cost-benefit model is simple and the cost and benefit functions may seem a bit arbitrary, the model shows the logic behind optimal protein levels and behind optimal parameter choices in general. An example concerns optimal plasmid copy numbers in yeast cells, which depend on tunable cellular costs and benefits of the plasmid. This artificial systems was called "genetic tug-of-war" because of the cost and benefit functions "pulling" the plasmid number towards lower or higher values (see Box 11.B).

Biology box 11.B Genetic tug-of war

Throughout the chapter, we mostly present simple examples involving the expression of one protein at a time. However, in nature, the situation is often more complex. To illustrate a system with two proteins, Moriya et al. [15] introduced an innovative method called "genetic tug-of-war". In this method, two genes were placed on a plasmid: one encoding a beneficial protein absent in the yeast strain (a leucine synthesis gene) and another encoding a target protein with an unknown benefit (a protein native to yeast). The plasmid was inserted into cells, which were then grown in a leucine-free medium for a few hours, allowing the cells to reach an optimal plasmid copy number.

A "tug-of-war" occurs because both proteins impose a cost, driving the optimal plasmid copy number down, while their (potential) benefits push it upward. Although each protein might have its own optimal copy number, both genes will have the same number of copies because they are on the same plasmid. As a result, the cell must reach a "compromise" where the combined costs and benefits of both proteins balance out. The leucine gene always provides a significant benefit, resulting in a high optimal copy number. In contrast, the benefit of the target gene varies and can even be zero, as these genes are already present in the yeast cell, and further overexpression may only increase costs without adding benefit.

So what did this experiments reveal about the proteins in question? By measuring the final plasmid copy number and comparing it to a control plasmid containing only the beneficial gene, the researchers could infer the effects of the target proteins. Low copy numbers suggested that wild-type levels were already optimal and that overexpression was harmful. Higher copy numbers indicated robustness to overexpression, while cases where the copy number exceeded the control highlighted a growth advantage provided by the target protein.



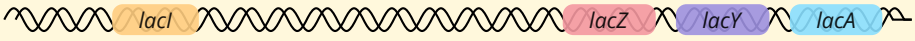
11.4 Cost and benefit of the lactose utilization system in *E. coli*

We now demonstrate the theoretical concepts from the previous sections with a canonical example for analyzing the cost and benefit of expressing a single gene, the Lac operon system in *E. coli* (see Box 11.C). The lactose utilization system, encoded by this operon, is a set of proteins that allows the cell to take up and metabolize lactose and is only expressed when lactose is present in the environment. In their pioneering paper from 2005, Dekel and Alon [12] performed a series of experiments to isolate the cost and benefit of genes in this operon, and to fit the missing parameters for their model.

For measuring cost, Dekel and Alon [12] designed an experiment where the benefit of the expression is zero, by growing the cells in defined glycerol medium lacking any lactose. Under these conditions, expressing the Lac operon

Biology box 11.C The Lac operon

The lactose utilization system and its regulation was the of first genetic regulatory mechanism to be studied and characterized, and won François Jacob and Jacques Monod the Nobel prize in 1965. Denoted the Lac operon, this section of *E. coli*'s genome contains three genes: *lacY*, *lacZ*, and *lacA*. Together, they encode the proteins that are responsible for the transport and metabolism of lactose. The Lac operon is preceded by another related one containing the *lacI* gene, which encodes a transcription factor called the *lac* repressor.



regulatory region Lac operon

(DNA double-helix not for scale)

In the absence of lactose, the *lac* repressor blocks the transcription of the Lac operon genes, and thus decreases the levels of the enzymes and transport proteins that enable the cell to consume lactose. This allows the cell to divert its resources away from these proteins as they do not contribute to growth and, therefore, fitness. When lactose appears in the environment of the cell, some fraction of it is converted to allolactose (by a side activity of β-galactosidase). Allolactose then binds to *lacI* and induces a conformational change that prevents it from binding DNA, alleviating the inhibition and initiating the transcription of the Lac operon genes.

has no benefit. Then, in order to alter the expression level, IPTG (isopropyl β-D-1-thiogalactopyranoside) was added at varying concentrations to the media. IPTG is a molecular mimic of allolactose, and similarly binds *lacI* to alleviate its repression of the Lac operon. However, unlike lactose and allolactose, IPTG cannot be metabolized by the cell and therefore provides no growth benefits. The growth rate of the cells in every condition was compared to the growth rate without IPTG, and the relative reduction in growth was defined as the cost (η). In order to quantify the relative *lac* expression , Dekel and Alon [12] used a standard assay for measuring β-galactosidase (lacZ) activity based on the optical readout of a colorful substrate, as well as SDS gel electropherosis. These results are shown in Figure 11.7A.

It was clear from the data that the cost is a non-linear function of the expression level. However, lacking a mechanistic model for this, a heuristic cost function was employed in order to fit the empirical data:

$$\eta(Z) = \eta^{\circ} \cdot \frac{\left(\frac{Z}{Z_{\mathrm{WT}}}\right)}{1 - \left(\frac{Z}{Z_{\mathrm{WT}}}\right)/M} \tag{11.3}$$

where Z is the expression level of the Lac proteins and Z_{WT} is their fully induced expression level in the wild-type strain. η° and M are fitting parameters and, as can be seen in Figure 11.7A, this cost function can be fitted well to the measured data using these values: $\eta^{\circ} = 2\% \pm 0.3\%$, $M = 1.8 \pm 0.3$. Interestingly, the cost for the unrepressed expression level of the wild-type cells was found to be $\eta(Z_{\mathrm{WT}}) = \eta^{\circ}/(1 - 1/M) \approx 4.5\%$.

To verify that this cost function remains relevant also after evolving the cells, Dekel and Alon [12] measured the lacZ expression and growth rate of strains that were evolved under constant concentrations of lactose (0.2 mM and 5 mM) for more than 400 generations. Indeed, the expression of the Lac operon changed, as well as the growth rate in the defined glycerol medium. However, the relationship between them remained the same (i.e. still roughly following Eq. (11.3)).

Next, Dekel and Alon [12] turned to quantify the benefit function, basing their model on the known transport and catabolism kinetics of the Lac system, i.e. assuming that the increase in growth rate (B) is proportional to the rate of lactose use, and follows the following function:

$$B(Z, L) = \delta \cdot \frac{Z}{Z_{\mathrm{WT}}} \cdot \frac{L}{K_Y + L} \tag{11.4}$$

where L is the concentration of lactose in the cell, K_Y is the Michaelis-Menten constant of LacY, Z is the LacZ protein level (i.e., as before, representing the Lac operon expression level), and δ is the relative growth advantage per LacZ

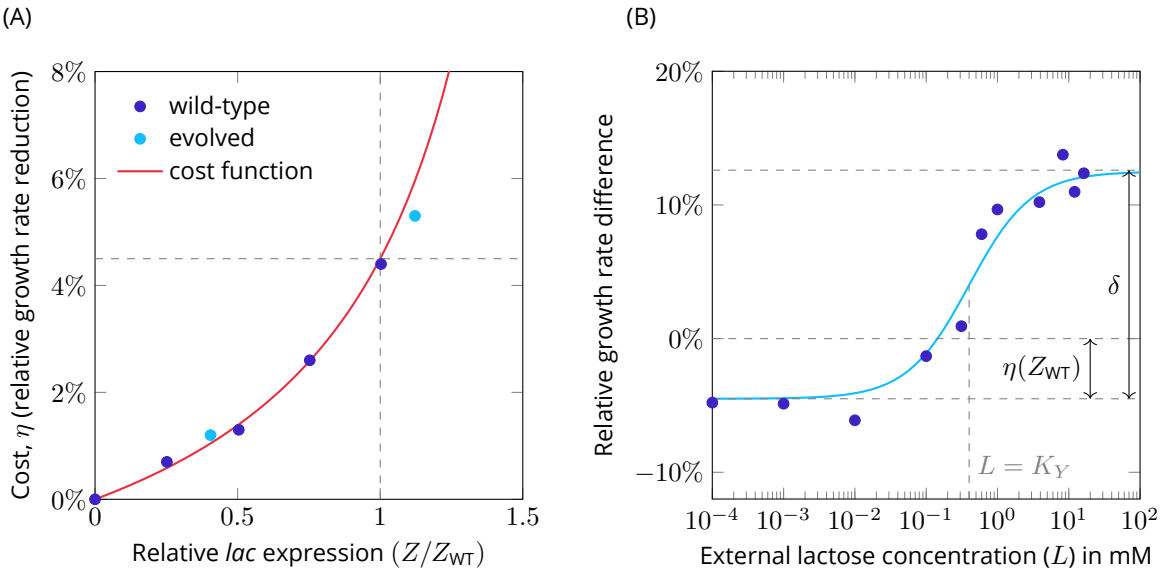


Figure 11.7: (A) Lac operon cost function Eq. (11.3), with parameters $\eta^\circ = 0.02, M = 1.8$ fitted to empirical data – Wild-type *E. coli* (dark blue points) was grown in defined glycerol medium with varying amounts of IPTG. The cost (η) is defined as the relative reduction in growth rate compared to no-IPTG conditions. The *x*-axis is the ratio of protein level to the fully induced wild-type protein level (Z/Z_{WT}). Also shown are the costs and Lac expression levels of strains evolved at 0.2 mM or 5 mM lactose (cyan points). (B) Growth rate effects of *lac* proteins as a function of lactose concentration – The data points (dark blue) represent cells growing at saturating levels of IPTG and varying concentrations of lactose. The growth rate difference (*y*-axis) is relative to cells growing without IPTG nor lactose. Here it is modeled as a benefit-cost difference. The value indicated as $\eta(Z_{\text{WT}})$ is the cost of the fully induced lac system at zero lactose, and δ is the benefit of lac induction at saturating lactose levels. The light blue line follows Eq. (11.5) with parameter values $K_Y = 0.4$ mM, $\delta = 17\%$, and $Z = Z_{\text{WT}}$. Note that K_Y represents the lactose concentration where LacY is half-saturated and therefore corresponds to the midpoint between the minimum and maximum of the benefit function, i.e. $\delta/2 - \eta(Z_{\text{WT}})$.

molecule at saturating lactose concentration. K_Y is known to be 0.4 mM from biochemical assays of LacY. Based on the kinetic parameters for both LacY and LacZ, when the Lac operon is fully induced and the lactose levels are saturating, growth rate should increase by 17% relative to conditions with no lactose (but when still fully expressing the Lac proteins). This means that we can use $\delta = 17\%$.

It is difficult to create conditions where a protein brings a benefit, but has no cost at all. So, in order to see if the benefit function can fit experimental data, media containing saturating IPTG levels (constant cost) and different concentrations of lactose (varying benefits) was used. Since IPTG was saturating, the added lactose would not increase the expression of the Lac operon any further. As before, the relative change in growth rate was measured and compared to the no-IPTG no-lactose condition. According to the cost/benefit model, the relative change should follow the following function:

$$g(Z,L) = B(Z,L) - \eta(Z) = \underbrace{\delta \cdot \frac{Z}{Z_{\text{WT}}} \cdot \frac{L}{K_Y + L}}_{\text{benefit}} - \underbrace{\eta^\circ \cdot \frac{\left(\frac{Z}{Z_{\text{WT}}}\right)}{1 - \left(\frac{Z}{Z_{\text{WT}}}\right)/M}}_{\text{cost}} . \tag{11.5}$$

Note that when there is close to no lactose in the medium, the relative growth rate will be a negative number since the cells grow slower due to the extra cost of expressing the Lac operon when IPTG is present, i.e. of the unrepressed expression level of the wild-type cells: $g(Z_{\text{WT}}, 0) = -\eta(Z_{\text{WT}}) \approx -4.5\%$. In general, plotting the measured data next to the predictions based on Eq. (11.5), we see that the fit is quite good – Figure 11.7B.

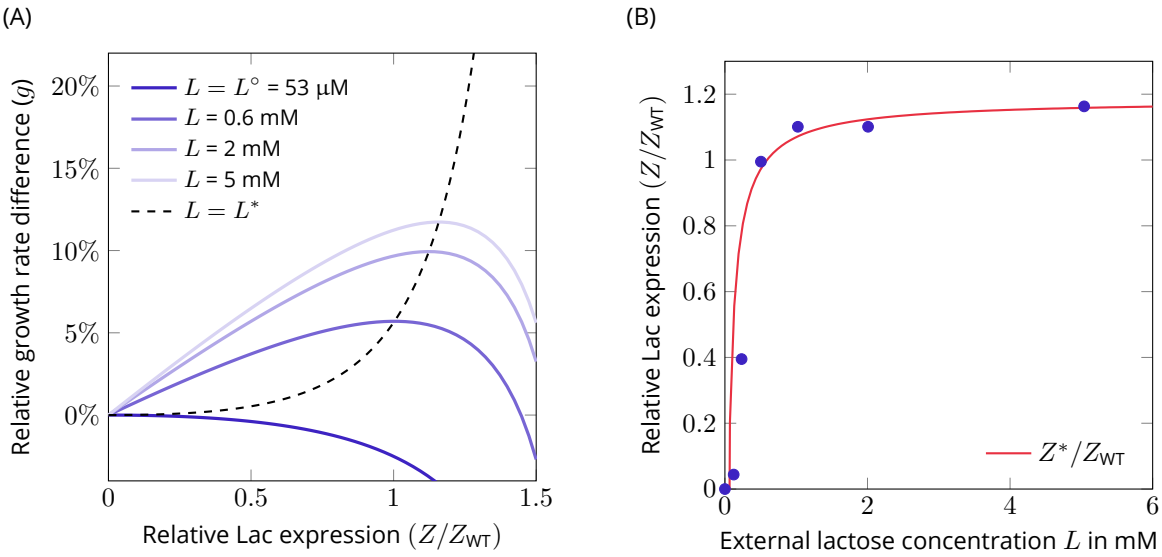


Figure 11.8: (A) Predicted relative growth rate of cells as a function of Lac protein expression – The relative growth rate difference, i.e. benefit minus cost, is shown for different concentrations of lactose (L), according to Eq. (11.5). The dashed black line represents the set of points with optimized expression for a range of lactose concentrations, following Eq. (11.6). The model predicts that over the course of evolution, the expression will be tuned to this value for every given lactose concentration. (B) Adapted LacZ activity of cells in serial dilution experiments – The evolved Lac operon expression level relative to wild-type cells is plotted here as a function of lactose concentration, L, present in the medium throughout the serial dilution experiment. The red line indicates the theoretical prediction of the optimal expression level (Eq. (11.6)).

Now, having a fully parameterized cost/benefit function, the optimal induction level (Z) can be expressed as a function of the lactose concentration (L), i.e. by finding the value of Z/Z_{WT} that satisfies $\partial g/\partial Z = 0$:

$$\begin{aligned} \frac{\partial g}{\partial Z} &= \frac{1}{Z_{WT}}\left[\delta\frac{L}{K_Y+L} - \eta^\circ\cdot\left(1-\left(\frac{Z}{Z_{WT}}\right)/M\right)^{-2}\right] \\ \left.\frac{\partial g}{\partial Z}\right|_{Z^*} = 0 \quad &\Rightarrow \quad \frac{Z^*}{Z_{WT}} = M\left(1-\sqrt{\frac{\eta^\circ}{\delta}}\cdot\sqrt{1+\frac{K_Y}{L}}\right). \end{aligned} \tag{11.6}$$

Note that, when $L \le L^\circ \equiv K_Y/(\delta/\eta^\circ - 1)$ then the value of Z^* in the above equation is negative. Therefore, the optimal *lac* expression level for lactose concentrations lower than the critical value, L°, is zero.

We can see in Figure 11.8A examples for the relative growth rate difference (g) as a function of Z for several levels of lactose, and what the optimal expression level is, based on the above solution. As expected, at the critical concentration of L° the function is monotonically decreasing. In other words, the external lactose concentration must be higher than L° in order for the benefit to outweigh the cost, otherwise the cells are better off ignoring the lactose.

We can further define L^* as the concentration of lactose that would be optimal for a given LacZ expression level Z. This can be easily done by solving Eq. (11.6) for L:

$$L^* = \frac{K_Y}{\frac{\delta}{\eta^\circ}\left(1-\left(\frac{Z}{Z_{WT}}\right)/M\right)^2 - 1} \tag{11.7}$$

and plugging in the values we obtained previously ($K_Y = 0.4$ mM, $\delta = 17\%$, $\eta^\circ = 2\%$ and $M = 1.8$) we get that $L^* \approx 0.6$ mM, and $L^\circ = 0.053$ mM. Indeed, as can be seen in Figure 11.8A, the optimum of the $L = 0.6$ mM line (dark blue) is reached approximately at $Z/Z_{WT} = 1$.

Finally, Dekel and Alon [12] tested whether cells growing for many generations in a constant environment with a specific concentration of lactose would evolve to optimize their expression of the Lac operon. The cells were grown in 10 mL cultures with a defined minimal medium, supplemented with lactose as the sole carbon source at varying

concentration. Whenever the cells reached a high optical density, they were diluted by a 100-fold into a new shake flask with the same medium. After many rounds of serial dilution, equivalent to 400-500 generations, the cells were isolated and the activity of LacZ was measured for each lineage. Figure 11.8B shows the final LacZ activity for each evolutionary condition (L). Also here, the predicted optimal expression based on Eq. (11.6) matches the measured data.

11.5 Proteins and growth: insights from constraint-based models

11.5.1 How do protein costs arise in a cell?

We saw how a protein's effects on cell growth can be described by effective cost and benefit functions. In a simple example, we assumed that the metabolic flux in a pathway contributes to cell growth, while enzyme expression contributes negatively, and we described these effects by simple cost and benefit functions. Similar costs and benefits can be defined based on experiments in which protein expression is controlled from the outside and the resulting changes in cell growth are measured. But what do these cost and benefit functions mean, and how do they arise from biochemical processes and economic compromises in cells?

Empirical protein cost functions describe overall effects quite well, but they do not tell us what is happening in a cell. First, they cannot describe how enzyme costs arise precisely and what are the proteomic rearrangements behind them. Second, they do not capture the detailed competition for resources (e.g. for individual amino acids or trace elements), and how this competition plays out between proteins of different size and composition. And third, assumping that costs and benefits can be clearly separated is a strong assumption. In reality, changes in protein expression have all kinds of effects, and it may be impossible to distinguish between benefits and costs at all – there will only be some overall effect on growth.

To better understand what "protein cost" even means, and how growth rate deficits come about, we may study cell-wide rearrangements of resources as described by whole-cell models. In such models, we can change protein levels at will and predict the effects on growth without assuming cost or benefit functions *a priori*. Studying such models can teach us how protein-specific and network-wide aspects of protein costs are related, how optimal protein levels depend on external conditions and biochemical details of cells, and how strongly the growth rate decreases when expression levels deviate from these optimal values.

11.5.2 Predicting the effects of protein levels on growth in whole-cell models

In the previous chapters, you saw cell models that–in principle–can predict protein levels that maximize cell growth. We can also use them to predict growth-rate effects of changes in protein levels. Let us have a look at some of these predictions! For simplicity, we consider linear resource allocation models, that is, models in which all rate laws are approximated by $v_l = e_l\, \kappa_l$ with constant enzyme efficiencies[1] κ_l. In such models, a protein has two direct effects. First, it occupies a part the protein budget (described, for example, by a density constraint in the model) and thereby decreases the budget for other proteins and, in particular, for ribosomes. Second, if the protein is a metabolic enzyme, increasing its amount will allow the cell to increase metabolic fluxes and therefore precursor production. Both effects are indirect and may involve rearrangements of the proteome and of metabolic fluxes. In a cell model, these opposite effects require a best compromise, which can be determine in models by maximizing the growth rate. By screening the possible enzyme levels, we obtain curves relating enzyme levels to growth rates as in Figure 11.1.

In practice, for constraint-based cell models the way to obtain such curves is always the same. A constraint-based model consists of a set of equalities and inequalities that define a set of solutions, each corresponding to a possible cell state with a certain cell growth rate and a certain expression level of the protein in question. By projecting this high-dimensional set onto a plane spanned by these two variables , we obtain a 2-dimensional set of possible protein/growth rate pairs. We may now assume that the cell, for each protein level, realizes the maximal possible growth

[1] In kinetic models (e.g. as described in chapter 6), the κ_l values would be variable and dependent on metabolite levels. This would change the results, but here we assume that linear models capture the most important effects, including rearrangements of the proteome.

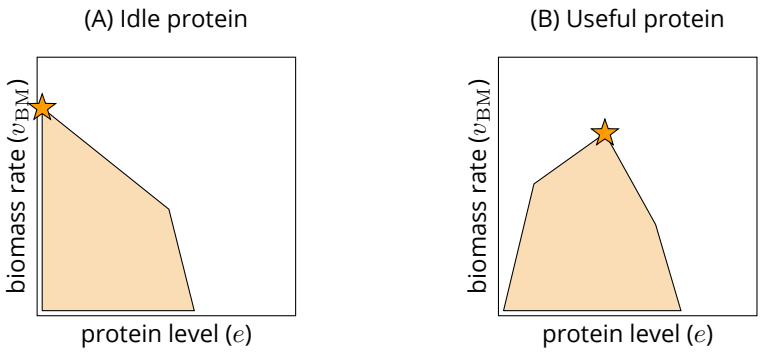


Figure 11.9: Protein/growth curves obtained from constraint-based metabolic models (schematic example) – A constraint-based metabolic model with fluxes and enzyme levels as variables defines a solution space, given by a high-dimensional polytope. By projecting the polytope onto a plane spanned by a protein level of interest and biomass rate, we obtain a feasible set in protein/biomass space. If the original problem is a Linear Programming (LP) problem, this set is a convex polygon. Its silhouette curve defines the biomass/enzyme curve. In models with a fixed total protein amount, the biomass rate per protein amount can serve as a proxy for the cell growth rate (see Chapter 7). The plots show, schematically, an idle protein (A) and a useful enzyme (B), where the optimum point is marked by an orange.

rate. The resulting cell states will lie on the upper boundary of this set (see Fig. 11.9), a simple prediction of the curves in Figure 11.3.

If our model is linear, the projected set is a convex polygon, with a bounding curve consisting of straight lines as shown in Figure 11.9. For an idle protein, the curve starts at a maximum at zero expression and decreases monotonically. For a useful protein, for example a metabolic enzyme used in the current conditions, the curve has a maximum at some positive expression level. In both cases, the decreasing part of the curve can be seen as a Pareto front describing a trade-off between our protein level and cell growth. In the case of a "useful protein", the curve has also an increasing branch on the left, describing a "win-win" situation in which higher enzyme levels allow for higher growth rates.

11.5.3 Minimal models trading metabolism against protein production

As a simple example, we consider a variant of the first cell model in Chapter 8 (Section 8.3). Metabolism and protein production are described by two overall reactions:

$$\mathrm{EXT} \xrightarrow{v_1(e_1)} \mathrm{PRE} \xrightarrow{v_2(e_2)} \mathrm{MAC} \tag{11.8}$$

The first reaction converts external metabolites (EXT) into precursors (PRE) and is catalyzed by an effective "metabolic enzyme" E_1 representing all metabolic enzymes in the cell. The second reaction converts the precursors into macromolecules (MAC) and is catalyzed by the "effective ribosome" E_2, representing ribosomes, chaperones, tRNA, RNA polymerases and other machines, which we assume to act in fixed proportions. For simplicity, we denote both types of machines as "enzymes", even if ribosomes are actually RNA-protein complexes; and for a first simple model, we assume that the catalyzed reaction rates v_i depend linearly on the enzyme levels ($v_1 = \kappa_1\, e_1$ and $v_2 = \kappa_2\, e_2$, with constant efficiencies κ_l). To relate the enzyme levels to the cell growth rate μ, we assume each of the reaction rates limits cell growth. This means that in optimal states the two rates must be equal ($v_1 = v_2$) because otherwise, one of them would be higher than necessary and protein would be wasted. Hence, a steady state is already guaranteed without any further constraints. Assuming a fixed protein budget, we put a bound $e_{\mathrm{tot}} = p_{\mathrm{tot}} - q$ on the sum of enzyme levels, which leads to a trade-off between e_1 and e_2. Here p_{tot} is the total protein budget and q is the size of the "Q sector" of other, non-modeled proteins in the cell. Altogether, our model contains the variables v_1, v_2, e_1, e_2, μ

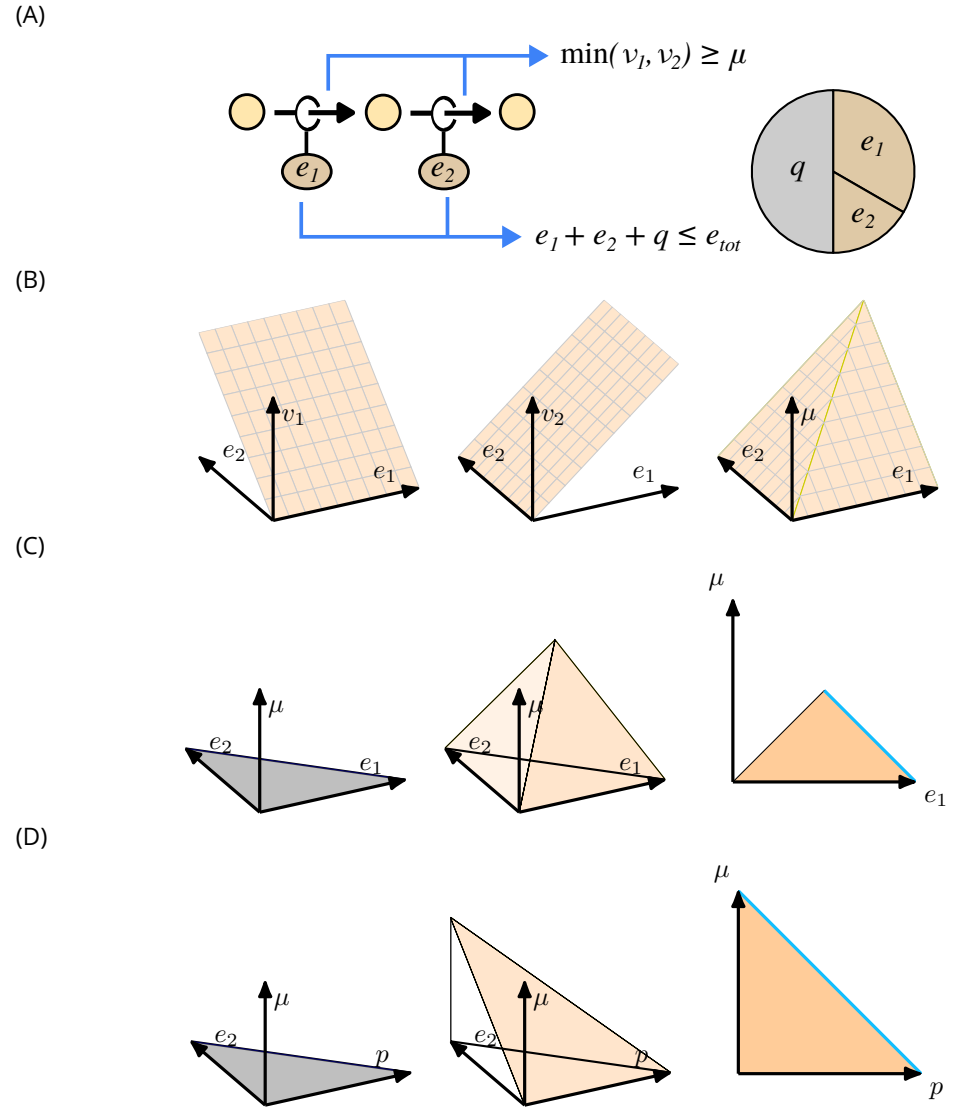


Figure 11.10: A minimal cell model can predict simple protein/growth curves – (A) Model with two reactions, catalyzed by two machines called "metabolic enzymes" and "effective ribosome". Each of the fluxes limits the cell growth rate, while the enzyme levels are limited by a fixed total protein budget. (B) According to Eq. (11.9), the reaction rates depend linearly on enzyme levels, and the growth rate μ is limited by the minimum of the two rates, so the achievable growth rate μ depends on e_1 and e_2. (C) Due to the protein bound, the solution space becomes a tetrahedron. Projecting it to the e_1/μ plane yields a triangle (right) whose boundary is a piecewise linear protein/growth line. The curve consists of two straight lines with a maximum point in between. (D) For an idle protein, adding the protein level p to the Q sector and thereby decreasing the available enzyme budget leads to a single decreasing line with an optimum at zero expression, and zero cell growth where the idle protein takes up the entire available protein budget. If protein level and cell growth are seen as simultaneous optimization objectives, the falling part of the boundaries in (C) and (D) (marked in blue) forms a Pareto front.

and the parameters κ_i, e_{tot}, and q. Its feasible states must satisfy the constraints

$$\begin{aligned} \mu &\le \min(v_1, v_2) \\ v_1 &= \kappa_1\, e_1 \\ v_2 &= \kappa_2\, e_2 \\ e_1 + e_2 + q &\le p_{\mathrm{tot}}. \end{aligned} \tag{11.9}$$

For each choice of parameter values ($p_{\mathrm{tot}}, q, \kappa_1, \kappa_2$), the constraints define a range of possible states $(\mu, v_1, v_2, e_1, e_2)$ (see Figure 11.10). The model (11.9) resembles the first model in Section 8.3 in Chapter 8 with two main differences. First, the cell growth rate is not directly given by the reaction rates, but only limited by them: ($\mu \leq v_1$ and $\mu \leq v_2$). Second, we do not explicitly require a steady state; but since non-steady states with $v_1 \neq v_2$ are suboptimal, the model predictions will be equivalent. As shown in Figure 11.10, the constraints (11.9) define a feasible polytope in the space spanned by e_1, e_2, and μ. The projection to the e_1/μ plane yields a feasible triangle with a single growth-optimal point (see Figure 11.10(B)).

What can we learn from this model? To simulate the expression of an idle protein, we assume that its abundance p adds to the Q sector, thus decreasing the remaining budget for e_1 and e_2 (straight line in Figure 11.10(D)). In our model (11.9), the model variables μ, v_1, v_2, e_1 and e_2 scale proportionally: therefore, a decrease of the available enzyme budget $e_1 + e_2 = p_{\mathrm{tot}} - q$ leads to a proportional decrease of the growth rate. With an idle protein adding to the Q sector, the growth rate changes as

$$\mu \sim \frac{e_{\mathrm{tot}}}{p_{\mathrm{tot}} - q} = \frac{p_{\mathrm{tot}} - q - p}{p_{\mathrm{tot}} - q} = 1 - \frac{p}{p_{\mathrm{tot}} - q}. \tag{11.10}$$

or

$$\mu = \mu_0 \left[1 - \frac{p}{p_{\mathrm{tot}} - q}\right]. \tag{11.11}$$

This means that the protein/cost curve is a falling straight line that hits 0 where $p = p_{\mathrm{tot}} - q$, that is, where the idle protein takes up all the available protein budget, and no budget is left for the enzymes. In summary, our model predicts that useful proteins that contribute to the "metabolic enzyme" pool have a growth curve with an optimum at a positive enzyme level. Idle proteins, in contrast, reduce the protein budget and therefore growth. The predictions assume a reallocation of protein resources between metabolism and translation, but they do not describe individual proteins with individual costs and benefits. To describe this, more detailed models are needed.

11.5.4 Flux-balance models with enzyme constraints

Our simple model (11.8) predicts simple protein/growth curves, but it cannot describe metabolic enzymes individually. However, we can easily expand it by replacing the "effective metabolic reaction" by a metabolic network. We obtain more detailed protein sector models, as in Constrained Allocation FBA (CAFBA) [16], an FBA model with enzyme constraints and a ribosome sector (see chapter 5). In such models, fluxes satisfy mass balance ($\mathbf{N}\,\mathbf{v} = 0$), catalytic constraints $v_i = \kappa_i\, e_i$, and maybe heuristic flux bounds (e.g. positivity constraints $v_i \geq 0$, after reorienting the reactions to impose realistic flux directions). We further assume a limited enzyme budget $\sum_l e_l \leq e_{\mathrm{tot}}$ and treat the biomass flux v_{BM} (or a similar flux objective $B = b_J \cdot v$) as the objective.

Putting all this together, we can compute growth-optimal states. How will changes in enzyme parameters, such as catalytic rate k_{app} or molecule size, change the predicted growth rate? In FBA with enzyme constraints, each enzyme has two direct effects. On the one hand it catalyzes a flux v; on the other hand it occupies a part of the protein budget. In optimal states, the effect on the flux must be beneficial (i.e. the flux v_i must have a positive marginal effect on the biomass rate, because otherwise the enzyme would not be expressed), and using the protein budget is costly (because of opportunity costs). We can now treat the enzyme level as a tunable parameter, optimize all other model variables, and compute the growth rate. By screening a range of enzyme levels, we obtain a parameter/growth curve, the upper line of the feasible set. At the optimal enzyme level, we obtain the optimal overall biomass/enzyme productivity–that is, the biomass production per enzyme–which can be translated into a predicted cell growth rate (see Chapter 7).

While FBA-like models are more detailed than our previous minimal model, the procedure for computing protein/-growth curves remains exactly the same. In the space of model variables, the model constraints define a polytope of valid states. Projecting this polytope onto a plane spanned by protein level and biomass rate leads to a feasible polygon. Its silhouette line describes the maximal possible biomass rate, satisfying all the model constraints, as a function of the protein level (see Figure 11.9).

Qualitatively, the lines look like the curves in Fig. 11.5 or in the previous simple model. For idle proteins (or enzymes that do not contribute to a cost-efficient metabolic strategy in the given conditions), the line is strictly decreasing, with a maximum at zero expression. For useful proteins, the line has a maximum at a finite (non-zero) expression level.

11.5.5 Complex whole-cell models

Resource Balance Analysis (RBA) models (see chapter 10) are even more realistic than FBA. RBA models can describe macromolecular processes in great detail. Instead of assuming a given biomass composition, they optimize it together with the metabolic state, requiring that metabolism supplies all the precursors for macromolecules, which in turn are needed to produce other macromolecules or catalyze metabolic reactions. Hence, metabolism and macromolecule synthesis depend on each other and form a big feedback loop. Mathematically, possible cell states can be described in a space with the growth rate as one dimension and all other cell variables (fluxes and macromolecule concentrations) as the other dimensions. For each given growth rate, these other variables must satisfy linear constraints based on physical laws. Therefore, the space of valid cell states consists of high-dimensional polytopes, "stacked" along an extra, continuous growth rate dimension.

To obtain protein/growth curves, we proceed as before: we project the set of feasible states (including the growth rate μ as one of the variables) onto a plane spanned by μ and the protein level in question. The main technical difference is that now we need to screen possible growth rate and determine, for each growth rate, a lower and upper bound for our protein level (where all other variables can be adjusted). However, just like before the silhouette line of the projected set yields the protein/growth rate curve (see Figure 11.11). Unlike in FBA, our model contains nonlinear dilution rates of macromolecules and, possibly, growth rate-dependent parameters, $v_{\mathrm{dil}} = \mu\, c_{\mathrm{macromol}}$, so the silhouette lines may be curved.

Similar to flux variability analysis in FBA (see chapter 5), we may run a "resource variability analysis" to explore the allowed region and to compute the Pareto front (Figure 11.11). At each given growth rate (shown in a normalized form, as a relative fitness), a machine or a single protein can show a range of possible expression levels. At the maximum growth rate, this range will usually shrink to a point. To interpret the plot, it is good to remember that there is no fundamental difference between "variability analysis" and "multi-objective optimization". If a machine concentration x and the cellular growth rate are seen as objectives to be maximized, the boundary line yields a protein/growth curve as in Figure 11.1, and the part on the right can be seen as a Pareto front. In constraint-based models, the "objectives" (such as growth rate) and "model variables" (such as protein amounts) are not fundamentally different types of variables, but part of a large set of variables that mutually constrain each other. "Multi-objective optimization" is just a way to describe these effective constraints, and to present it as a trade-off between the two variables.

So if we aim at predicting protein/growth curves, what do we gain by using RBA models? Like in FBA models, producing a useless protein will make the growth rate decrease, while for useful proteins, any deviation from its optimal, positive amount decreases the fitness. However, compared to FBA, the predictions account for more subtle effects and more complex adaptations of the cell, including changes in biomass composition. Unlike in FBA, varying protein levels will change the demands for biomass precursors, and therefore the required metabolic fluxes–which in turn lead to changing enzyme demands! For example, using proteins that contain trace elements such as metal ions will increase the demand for these ions. In contrast, when these ions are scarce, this has two different effects. First, it leads to a reallocation of protein towards pathways that scavenge these trace elements. But in addition, if ions are costly to obtain, the expression levels of the ion-containing proteins may be decreased. The logic may also hold for nitrogen. Under low-nitrogen conditions in the environment, a simulated cell may both increase nitrogen import and decrease the use of proteins with a high nitrogen content. Hence, metabolic production and protein usage are tightly entangled via biochemical processes and cellular economics, and all this is reflected in the predicted protein/growth curves.

11.6 Protein and growth: insights from mechanistic models

In this section, we stop taking for granted some of our previously made assumptions – most prominently, the notion that protein expression is optimal due to having been shaped by evolution – and consider the regulatory mechanisms actually making the cell behave according to them.

Why should we now start questioning our notions? All models used by us throughout this chapter to understand the effects of protein expression rely on constraints, which define the set of possible cell states and thus the set of cell behaviors realistically achievable in real life. However, not all constraints are of the same nature. Some stem from the first principles of molecular biology, such as the non-negativity of concentrations or mass conservation constraints

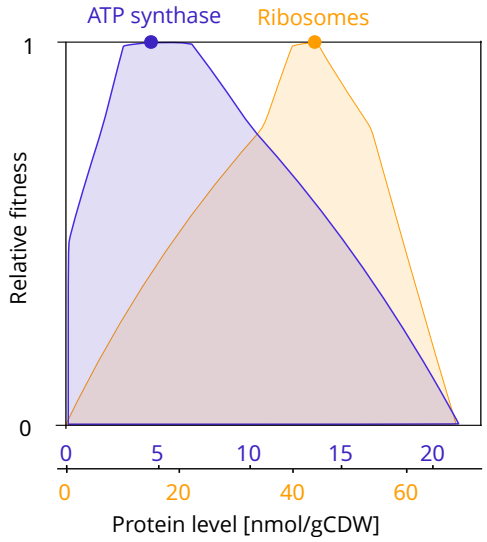


Figure 11.11: Protein expression and growth rate effects predicted by a Resource Balance Analysis (RBA) model – Protein/growth curves for ATP synthase and ribosomes were computed by resource variability analysis in a genome-scale RBA model of *B. subtilis* bacteria [17]. Cell growth, normalized to the maximal possible growth rate (and called "relative fitness") is shown on the x-axis and protein abundances are shown on the y-axis. The feasible sets were computed via resource variability analysis, similar to flux variability analysis in FBA (see chapter 5) and indicate the achievable minimal and maximal values of ribosome and ATP synthase expression, assuming possible rearrangements of all cellular variables. The maximal ribosome amount decreases at higher growth rates, indicating a trade-off between ribosome amount and cell growth (similar for ATP synthase). Figure redrawn from [18].

in FBA and RBA. Others, rather than addressing the question of what rules the cell **must** obey, focus on how the cell **tends** to behave. Such constraints are called "phenomenological" since they capture commonly observed phenomena without a fundamental explanation. An instance of this is the assumption of linear correlation between the cell's growth rate and its overall ribosome abundance (protein synthesis budget) often made in minimal constraint-based models [19], which reflects the trends observed in experimental data. In this vein, optimality may be considered the ultimate phenomenological constraint, with frameworks like FBA and RBA relying on it to predict the cell's state simply because growth rate maximization is often observed in nature.

However, in some cases, these phenomenological constraints may be invalid, as no inherent properties of the cell prevent their violation. Looking at the effects of protein expression in particular, at high protein production rates, the relationship between the 'useless' protein mass fraction and the growth rate deviates from the straight line [20] predicted by minimal constraint-based models [19] (e.g. the upper boundary in Figure 11.12D) . Moreover, the regulation of certain metabolic proteins' expression in *E. coli* has been experimentally determined to be suboptimal [21]. Suboptimal scenarios may also arise – and be particularly common – when we engineer cells with synthetic proteins, as the cell, by definition, has not had time to evolve and optimize their expression.

Since phenomenological constraints can sometimes be misleading, instead of enforcing commonly observed cell behaviors (which often means best-possible behaviors), it can be useful to ask: "what are the actual cellular regulation mechanisms enabling these phenomena?" To tackle this question, we can use mechanistic ordinary differential equation (ODE) models, which capture the dynamics of different cellular variables. These models still use 'first-principle' constraints (like the laws of enzyme kinetics), but bake them into the form of the differential equations rather than enforce them explicitly. The equations also incorporate the terms for particular molecular regulation mechanisms believed to make the cell closely (but possibly not perfectly) adhere to the phenomenological constraints. Due to the complexity of living systems, such ODE models are usually coarse-grained, with each variable describing the average dynamics of multiple biomolecules with similar functions and properties. The degree of this coarse-graining can be varied based on the desired level of detail and accuracy, as well as the scenario considered.

In many cases, mechanistic models reproduce constraint-based modeling predictions, either by simulation or analytically. The latter involves solving a model for the steady state (i.e. setting all ODEs to zero) and rearranging the terms of the equations. For example, models based on the Flux-Parity Regulation theory [20, 22], which capture the regulation of ribosomes by the ppGpp signaling molecule, whose abundance is proportional to the ratio between charged and uncharged tRNA concentrations, algebraically yield the same linear relation between useless protein expression and its cost to the cell (see Box 11.D and Figure 11.12).

Box 11.D Deriving the cost of useless protein expression

To derive a mathematical expression for the cost of expressing a protein without any beneficial influence on cell growth, we start by defining a simple mechanistic *E. coli* cell model from first principles, which is given by Ordinary Differential Equations (11.12) and proteome allocation relations from Equations (11.13). The model captures the total protein biomass of a growing cell population (M) and the uncharged and aminoacylated tRNA concentrations per cell (T_u and T_c, respectively). The fractions of the cell's proteome taken up by the ribosomal proteins, carbon metabolism proteins, housekeeping proteins (native proteins not in the two previous sectors) and the useless protein are respectively denoted as ϕ_r, ϕ_a, ϕ_q and ϕ_x. The ODEs below reflect the following modeling assumptions, which have been informed by experimental studies [22, 20]:

- The total protein mass density per unit cell volume, M, is constant, so the rate of the tRNA species' dilution due to cell growth (i.e. volume expansion) equals the rate of biomass increase λ. This also means that the protein mass fractions are proportional to protein concentrations per unit cell volume.
- tRNA aminoacylation, consuming uncharged and producing charged tRNAs, is catalyzed by metabolic proteins (hence its dependence on ϕ_a). Similarly, translation (and thus protein biomass synthesis) is catalyzed by ribosomes and depends on the abundance of aminoacyl-tRNAs. Each reaction's rate exhibits a Michaelis-Menten dependence on the concentration of its precursor.
- The signaling molecule ppGpp, whose level is given by the ratio of charged and uncharged tRNA abundances, controls proteome allocation between ribosomal and metabolic gene. This is captured by a Hill function.
- The rate of RNA transcription, including that for tRNAs, is proportional to the cell's growth rate.

$$\begin{aligned}
\frac{dM}{dt} &= \frac{\gamma T_\mathrm{c}}{K_\mathrm{trans} + T_\mathrm{c}}\phi_\mathrm{r} M = \lambda M \\
\frac{dT_\mathrm{c}}{dt} &= \frac{\nu T_\mathrm{u}}{K_\mathrm{aa} + T_\mathrm{u}}\phi_\mathrm{a} - \frac{\gamma T_\mathrm{c}}{K_\mathrm{trans} + T_\mathrm{c}}\phi_\mathrm{r} - \lambda T_\mathrm{c} \\
\frac{dT_\mathrm{u}}{dt} &= \frac{\psi(T_\mathrm{c}/T_\mathrm{u})}{K_\mathrm{ppGpp} + T_\mathrm{c}/T_\mathrm{u}} \cdot \lambda - \frac{\nu T_\mathrm{u}}{K_\mathrm{aa} + T_\mathrm{u}}\phi_\mathrm{a} + \frac{\gamma T_\mathrm{c}}{K_\mathrm{trans} + T_\mathrm{c}}\phi_\mathrm{r} - \lambda T_\mathrm{u}
\end{aligned} \tag{11.12}$$

where $\phi_\mathrm{q} = const,\ \phi_\mathrm{x} = const,$

$$\phi_\mathrm{r} = (1 - \phi_\mathrm{q} - \phi_\mathrm{x}) \cdot \frac{T_\mathrm{c}/T_\mathrm{u}}{K_\mathrm{ppGpp} + T_\mathrm{c}/T_\mathrm{u}},\ \phi_\mathrm{a} = (1 - \phi_\mathrm{q} - \phi_\mathrm{x}) \cdot \frac{K_\mathrm{ppGpp}}{K_\mathrm{ppGpp} + T_\mathrm{c}/T_\mathrm{u}} \tag{11.13}$$

For a cell in the steady state, tRNA concentrations are in equilibrium. Taking $\frac{dT_\mathrm{c}}{dt} = 0$ and $\frac{dT_\mathrm{u}}{dt} = 0$ and substituting the definitions from Equation (11.13), we can observe that all ϕ_x-dependent terms cancel out. Hence, charged and uncharged tRNA abundances in the steady state are predicted to be unaffected by the useless protein's expression. Then, the steady-state cost of having the useless protein occupy the fraction ϕ'_x of the cell's proteome can be found according to Equation (11.14), where we use the fact that the ϕ_x-independent tRNA abundances T_u and T_c are the identical in the numerator and the denominator of the fraction.

$$\begin{aligned}
\eta(\phi'_\mathrm{x}) = \frac{\lambda(\phi_\mathrm{x} = 0) - \lambda(\phi_\mathrm{x} = \phi'_\mathrm{x})}{\lambda(\phi_\mathrm{x} = 0)} &= 1 - \frac{\frac{\gamma T_\mathrm{c}}{K_\mathrm{trans}+T_\mathrm{c}} \cdot (1 - \phi_\mathrm{q} - \phi'_\mathrm{x}) \cdot \frac{T_\mathrm{c}/T_\mathrm{u}}{K_\mathrm{ppGpp}+T_\mathrm{c}/T_\mathrm{u}}}{\frac{\gamma T_\mathrm{c}}{K_\mathrm{trans}+T_\mathrm{c}} \cdot (1 - \phi_\mathrm{q} - 0) \cdot \frac{T_\mathrm{c}/T_\mathrm{u}}{K_\mathrm{ppGpp}+T_\mathrm{c}/T_\mathrm{u}}} \Leftrightarrow \\
\Leftrightarrow \eta(\phi'_\mathrm{x}) = 1 - \frac{1 - \phi_\mathrm{q} - \phi'_\mathrm{x}}{1 - \phi_\mathrm{q} - 0} &= \frac{\phi'_\mathrm{x}}{1 - \phi_\mathrm{q}}
\end{aligned} \tag{11.14}$$

This linear increase in costs as the useless protein's abundance in the cell increases is likewise predicted by the constraint-based model in Section 11.5.3. In the present case, however, we made no assumption of optimality but rather considered the experimentally determined growth regulation mechanisms via ppGpp signaling. This relation is also concordant with experimental data (see Figure 11.12) for sufficiently low ϕ_x values. Greater useless protein abundances, which yield a significant divergence from this law, are theorized to activate the cellular stress response mechanisms neglected by the simple cell model in Equations (11.12)–(11.13) [22].

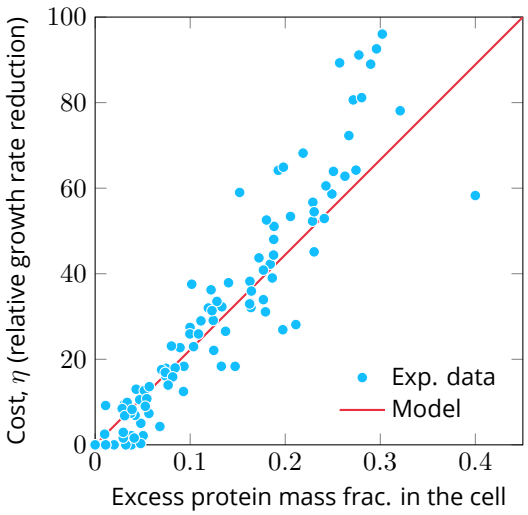


Figure 11.12: Experimentally measured synthetic protein mass fractions in the cell plotted against cell growth rates relative to cells expressing no useless protein (data taken from [20]). The red line represents the linear dependency observed in Figure 11.10 and derived from a mechanistic model in Box 11.D. At higher mass fractions, this linear relation is in practice violated, with costs growing superlinearly, due to the phenomena neglected by the mechanistic model.

In more detailed models, besides drawing from the ribosomal budget, protein expression is assumed to have positive and negative effects on other aspects of cellular functioning, such as its ATP and amino acid consumption and synthesis, all captured by different parameters. Varying these parameters individually and in conjunction yields different curve shapes, which can be compared to experimentally observed trends. This may help to deduce how exactly a given protein favors and hinders growth by looking at this protein's cost-benefit curves [23].

What if predictions from an ODE model disagree with reality, in the same way that constraint-based models' phenomenological constraints are sometimes violated? As discussed in Chapter 12, such observations provide valuable information, revealing that there is a regulatory mechanism or a biomolecular reaction currently overlooked by us. Namely, for the cost of expressing the LacZ or ΔtufB proteins, experimental measurements diverged significantly from the predictions obtained by simulating a mechanistic cell model from [23]. This discrepancy, however, was removed by redefining the protein degradation rate, changing it from a simple linear relationship to a Hill formula. The non-linearity was explained by noting that bacterial proteases primarily degrade misfolded proteins; hence, the need to model the dynamics of protein folding and its dependence on the overall protein expression levels.

11.7 Concluding remarks

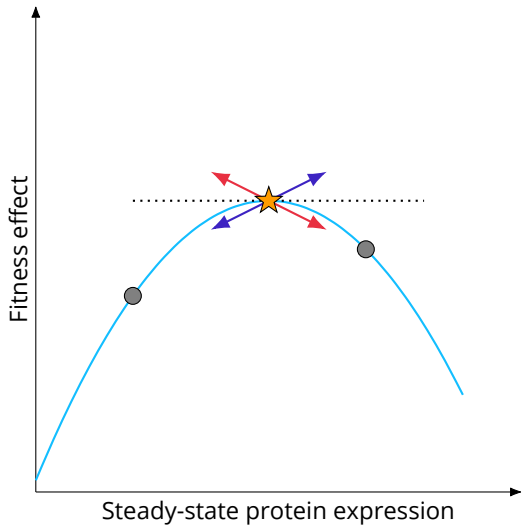
Any process in the cell can be viewed through the lens of its effect on the cell's fitness under given environmental conditions. By "fitness" we usually mean the efficiency with which the cell can increase its biomass – that is, the rate of cell growth. Furthermore, growth can easily be related to fluxes of matter in metabolic models we often use for predictions, . Considering growth rates also facilitates the economic analogy between a company's drive to increase its market value and the cell's incentive to build up biomass. This is the view we have adopted in this chapter; however, considerations would be similar for other proxy variables capturing a cell's fitness. A process of prime importance for cell growth is protein expression, since proteins enable the majority of metabolic reactions and comprise the greatest share of the cell's biomass, while their synthesis consumes most of the cellular energy (i.e. ATP molecules) and other resources. In this chapter, therefore, we discussed the question: how to determine the fitness effects of expressing a given protein? This consideration is crucial for predicting the cell's metabolic regulation and understanding how it has evolved, as well as for engineering the expression of heterologous proteins in cells to be optimal.

As we saw in this chapter, it is useful to separately consider the protein's positive and negative contributions to the growth rate – that is, its benefit and cost to the cell – and represent the protein's overall fitness effects as the difference of these terms (see Figure 11.13). This (artificial, rather than inherently physical) distinction allows one to

(A)

protein fitness effect F = fitness benefit B − fitness cost H

(B)



Expression usually assumed optimal:

↙ marginal benefit ↗
and
↖ marginal cost ↘
cancel out

Likely causes of non-optimality:

protein is non-native

cell is a non-model organism

protein function depends on other unconsidered proteins

Figure 11.13: The fitness effects of expressing a protein are (A) calculated as the difference between the protein's fitness and cost and (B) are usually assumed to be driven to an optimum by evolution.

draw parallels with economics, in which any activity's ultimate profitability for a company is calculated by subtracting the costs of undertaking it from its contribution to the revenue. The possibility of defining the overall effect as a linear difference is a simplifying assumption we make; however, it is justifiable by taking a linear approximation when operating with small variations in gene expression and is a natural property of certain metabolic models, such as those in which the cell's overall protein abundance is fixed. In other cases, a linear relationship can be obtained by redefining the fitness function, e.g. taking a logarithm of the ratio between the metabolic flux facilitated by the enzyme (its benefit) and its total amount (proportional to its cost) yields a difference between log-terms. Another convenient assumption is that the protein's abundance is the sole argument of the cost and benefit functions, which makes the effects of disparate proteins comparable to each other.

A protein's costs and benefit may be understood as purely empirical values relating the observed protein levels, varied by the experimentalist, to the cell's growth rates as measured using techniques discussed in the appendix. This approach is illustrated by the Lac operon case study, where first the cost is established in an experiment in which the protein of interest is rendered useless to the cell (by not having the enzyme's substrate present). Afterward, the benefit is determined by observing changes in cell growth for fixed protein expression levels with known costs. However, costs and benefits can also be treated more mechanistically by considering how they arise. Namely, benefits stem from proteins carrying out their biological functions. Meanwhile, expressing a given protein has direct "opportunity" costs – that is, different resources required for its synthesis and the space it occupies in the cell's proteome are unavailable for other proteins that are potentially necessary for growth. Moreover, changes in a protein's expression bring about global changes in the highly interconnected cellular metabolic network. Most prominently, increasing a protein's production raises demand for the cell's expression machinery, and increasing ribosome expression to meet this demand will itself consume more resources. Such effects may be called indirect costs, or "baggage". Defining mathematical models with different cost and benefit relations built in – potentially dependent on a given protein's molecular properties and enzyme efficiency – and comparing their predictions with experimental data can elucidate the fitness effects of expressing the protein.

Appreciating the costs and benefits of protein expression can help us understand how the cell manages its proteins and the evolutionary pressures that have shaped this behavior. If we assume that evolution drives a protein's expression towards enabling optimal fitness, its level can be predicted by retrieving a "sweet spot" between its cost and benefit. At this point, marginal (i.e. per one additional unit of protein produced) costs and benefits must cancel each other out, so that neither lowering nor increasing expression can improve fitness. Since costs per unit protein always increase, the existence of such an optimal level is conditional on the marginal benefits being positive. At the same time, to have an optimal protein level which is finite, protein expression must be constrained or must bring about diminishing returns as it increases further and further.

Nonetheless, care must be taken when applying the cost-benefit paradigm described in this chapter. Namely, optimization of gene expression for maximum fitness is merely an assumption backed by the fact that fastest-reproducing cells are usually favored by evolution. Hence, some proteins' expression may be near-optimal in some conditions but still be found away from the supposed "sweet spot" in others. To explain this, one can consider models which explicitly incorporate the mechanisms of gene regulation, as well as protein expression and processing, noting which modeling assumptions produce good agreement with experimental data.

Non-optimality may be particularly likely when the cell is engineered with synthetic genes encoding new proteins whose expression, unlike that of the cell's natural genes, has not yet been optimized by evolution. However, if proteins' fitness effects are studied and incorporated into models, their predictions can help to achieve such synthetic protein expression that cell population growth or production of desired compounds are maximized [22, 24].

An important caveat in all the above considerations is that costs and benefits are highly context-dependent. For instance, while almost any increase in a protein's expression in *E. coli* (which we focus on here) involves an observable cost, the same may not always be true for other organisms, such as *S. cerevisiae* [5, 6] or *B. subtilis* [17]. This is likely to stem from the fact that, depending on the conditions, some cells do not optimize their protein expression for maximum instantaneous growth, but rather maintain spare protein synthesis capacity to ensure adaptability to changing circumstances [25, 26, 18]. Within the cell, context-dependence may arise if the protein of interest operates in a complex with other proteins or is part of a metabolic pathway. Namely, for the latter, the expressed protein's fitness effects depend not only on substrate concentrations, but also on the levels of other proteins in the same pathway, which may decide whether upregulating the protein has positive marginal benefits by boosting the biomass flux or whether the entire pathway is too wasteful and the protein always has negative fitness effects.

Recommended readings

- **Empirical cost and benefit functions and evolution towards predicted optimal expression.** Erez Dekel and Uri Alon. Optimality and evolutionary tuning of the expression level of a protein. Nature, 436(7050):588–592, 2005. doi: 10.1038/nature03842.
- **A comprehensive description of metabolic control and optimality on the level of entire cells.** Frank J. Bruggeman, Maaike Remeijer, Maarten Droste, Luis Salinas, Meike Wortel, Robert Planqué, Herbert M. Sauro, Bas Teusink, and Hans V. Westerhoff. Whole-cell metabolic control analysis. BioSystems, 234:105067, 2023. doi: 10.1016/j.biosystems.2023.105067

Problems

Problem 11.1 Cost-benefit model for a metabolic enzyme

Maximize the fitness function Eq. (11.1) with a benefit function given by the pathway flux $B(e) = v(e) = v_{\max} \frac{e}{e+a}$ with constant parameters $v_{\max}$ and a, and without a prefactor b_J and an enzyme level $e \geq 0$. Show that your solution is a maximum (not just an extremum). Under what conditions (that is, for what parameter choices) will the enzyme be expressed (that is, have a non-zero expression level)?

Problem 11.2 Smooth enzyme-growth functions

The constraint-based models discussed in Section 11.5 predict piecewise linear protein-growth curves. Explain (in words) why in reality, the corresponding curves will be smooth instead of being piecewise linear.

Appendix sections

11.8 Measuring the fitness effects of protein expression

In order to quantify the fitness effects of protein expression, several different techniques or their combinations can be used. This is done by either measuring the growth rates and protein expression levels in the same method, or separately on the same cell culture. As expected, the growth rate measurement methods used vary depending on the cell type or the organism in question, and the protein quantity measurement approach depends on the specific protein of interest. Here, we will briefly review the most commonly used methods in literature.

11.8.1 Growth rates

Measuring growth rates of cells takes the form of measuring the number of cells or the amount of biomass at different time-points of a growing culture. Samples taken at these time-points are quantified for cell numbers, total cell volume, or biomass, either directly or those values estimating indirectly.

The most ***direct cell counting method*** for growth rate determination is **live microscopy**, where a dilute culture of growing cells is placed in a soft-agar pad or a microfluidics chip for direct observation and recording under a microscope. The snapshot images can later be processed using automated image analysis to obtain growth in cell size and cell numbers over time.

For direct counting of viable cell numbers, a widely used technique is the counting of **colony forming units** (CFU), which involves spreading a fixed volume of the culture on agar plates, and counting the number of colonies after a period of incubation. Although technically simple, it is a labour-intensive and time-consuming method, requiring incubation periods ranging from hours to days, depending on the organism.

More high-throughput methods of cell counting often involve passing a diluted suspension of cells through a counting chamber that counts the passing particles based on some physical property: the ability to generate a resistive pulse (**Coulter counter**) or light scattering (**flow cytometer**). Events recorded over time can then be used to count the number of cells in a fixed volume of the culture. These methods usually count all cells (dead or alive), but can be adapted with suitable dyes to only count living cells.

Among the indirect cell counting methods, the most commonly used is **turbidimetry** that involves using a spectrophotometer to measure the optical density (OD) of a microbial cell culture at a specific wavelength, typically 600 nm (OD_{600}), as the light passes through a unit path length (1 cm) of the suspended cells. The more the number of cells in a given volume, the more they scatter light, resulting in increased turbidity that can be measured as OD. By measuring OD of a growing culture at several time-points the rate of growth of the cells can be determined. Measurements can be automated using a 96-well or 384-well **plate reader** that records OD at regular programmed intervals. Although OD is often used as an indirect measure of cell density (numbers in a unit volume), it better estimates the total cell volume in a given volume of cell culture.

Cell biomass is another common method used to determine microbial growth. As the name indicates, this method involves concentration of the cell mass in a known volume of culture, and measuring its weight. In many applications, the concentrated mass is dried before weighing to obtain the **dry cell weight**. By taking cell biomass from cultures at different time-points the growth rate of the culture can be determined. This method can be less sensitive for detecting small changes. Therefore, it is only preferred for cells (like filamentous fungi) that don't stay uniformly resuspended in culture.

For determining growth rates of several different types of cells in a mixed culture, **pooled sequencing** based growth assays have recently emerged. These involve collecting samples of the mixed culture at different time-points and freezing them down. DNA from the frozen cells can then be sequenced to determine the amount of DNA of each cell-type to obtain relative cell numbers. Using spike-in controls, the relative numbers can be converted to their absolute

equivalents. The sequence determination is done using microarray chips or next-generation sequencing (NGS). By tracking the change in the cell numbers over time, growth rates of individual cell-types in a co-culture can be determined.

11.8.2 Protein expression

Several different methods can be employed to measure protein expression in cells. Most of these methods quantify the **protein amount**, rather than its rate of expression. However, if the methods are sufficiently high throughput, time-course snapshots of the protein amounts can help us rebuild the expression kinetics. Quantification methods also vary in whether they measure the total protein amounts, or the amounts of a specific protein in the cell.

Antibody-based methods are probably the oldest of protein quantification methods that measure the amounts of a specific protein. **ELISA** (enzyme linked immunosorbent assay) and **Western blotting** are the most common examples. These rely on the ability of antibodies to specifically bind the protein of interest and produce an output proportional to the amount of protein. The type of output produced may vary depending on the antibody used: colour, luminescence, or fluorescence, and the methods have to be carefully calibrated to get the protein numbers required.

Spectroscopic and fluorescence quantification methods are some of the easiest methods for protein quantification, and are therefore used widely. Bulk estimates of protein amounts can be made by measuring absorbance in the UV spectrum (280 nm wavelength), or chemical treatment of proteins to convert their amounts to a coloured output (for example, folin phenol to obtain blue colour). Similar approaches can be used to quantify specific proteins by **activity assays** for enzymes that are capable of converting colourless compounds to coloured ones (for example, LacZ converts ONPG to a yellow product), which can then be measured spectroscopically. If the protein of interest is naturally fluorescent (like GFP), its amounts can be determined by directly measuring fluorescence in a **plate reader** or a **flow cytometer**. Alternatively, non-fluorescent proteins can be fused with fluorescent protein tags for quantification. Similarly, if the protein of interest is luminescent (like luciferase), their amounts can be estimated by luminescence. The advantage of a plate reader is that kinetic measurements of OD and fluorescence at fixed time intervals can be used to measure the rates of protein expression and cell growth in the same experiment.

Mass spectrometry (MS) based methods are used to quantify amounts of all proteins in a sample by counting the number of reads of short peptide signatures derived from their lysis products. MS is a highly specialised and high-throughput method that normally quantifies the relative amounts of all proteins in a sample, but can be adapted for absolute quantification.

Ribo-seq (or ribosome profiling) is a relatively new high-throughput method that can be used to measure protein amounts in a cell by counting the number of ribosome-protected mRNA fragments in a sample. It relies on the observation that the amount of a protein in a cell is well correlated with the number of ribosomes translating its mRNA. The method can be used to observe relative changes in protein expression, or it can be calibrated to obtain absolute amounts.

Fluorescence microscopy is used to quantify the expression of fluorescent proteins in cells (or other proteins fused to them). This can be done by analysing fixed samples taken at different time-points, or by tracking per cell fluorescence in live microscopy. Together with tracking of cell division, this method can be used to measure growth rate and protein expression together.

Like fluorescence microscopy, **Flow-seq** is a method that can be used to track both the protein expression and growth rates of cells together, typically used for **pools of populations**. Cells are first sorted into several fluorescence bins using FACS, and the sub-populations sequenced to determine the number of cells of each type in each bin. The data can then be processed to obtain the fluorescence levels from individual cell types. If coupled with samples collected at different times, growth rates can also be calculated from changes in the population composition.

Pulsed-metabolic labelling of proteins is probably the most high-resolution method for determining the synthesis rates of proteins, rather than protein amounts. It involves adding a labelled amino acid to growing cells at a given time-point and then measuring the rate of incorporation of the amino acid to translating peptide chains in the cell by mass spectrometry of timed samples.

11.8.3 RNA expression

The amounts of expressed proteins in a cell, and the mRNAs encoding for the protein are well correlated, at least for native genes. This is due to a combination of the protecting effect of translating ribosomes on a transcript, thus increasing the half-lives of translation mRNAs, and the RNAP and ribosome homeostasis operating in the cell. Therefore, the amounts of mRNA of a given protein can often be taken as a proxy for the protein amounts.

Quantitative RT-PCR is the gold standard for measuring the relative or absolute amounts of a specific RNA sequence. It involves reverse transcription of extracted cellular RNA to DNA, and then using a quantitative PCR to measure the amounts of the DNA obtained. By taking appropriate cellular and spike-in controls, RNA amounts can be quite precisely quantified.

RNAseq is now a widely used technique to measure the amounts of all RNAs in a cell. It involves RNA to DNA reverse transcription, but the DNA is later sequenced using NGS sequencing. The number of reads obtained can then be used to estimate the expression levels of all mRNAs.

11.8.4 Controlling expression levels

Varying the level of transcription

At the transcription step, gene expression can be tuned by using promoter sequences with different regulatory mechanisms and transcriptional strengths. These are controlled by the binding of various cellular transcription factors at or near the promoter sequence in order to attract the RNA polymerase (RNAP) for transcription initiation. Depending on the organism, the RNAP type in question, and the kind of regulation operating on the promoter, its length can range from a few nucleotides (18 nt for T7 promoter) to over thousand nucleotides [27]. For genes required to be expressed at all times, “constitutive” promoters are used. For genes whose expression needs to be regulated in response to an inducer, promoters “inducible” by a small molecule, or another external input, may be used. Most commonly, promoters of the desired strength and induction properties can be identified from a library of previously characterized promoter libraries <ref>. However, it has recently been possible to predictively design synthetic promoter sequences with a user-defined transcription strength for commonly used organisms in biotechnology [28, 29]. The ability to regulate transcription of multiple genes using different promoter strengths can be used to optimize the expression levels of multiple enzymes in a metabolic pathway [30]. As already seen in the example above, overexpression by increasing transcription can cause burden in the cell as the protein expression from the transcribed mRNA takes up cellular resources.

Varying the level of translation

As with promoter-based transcriptional control, at the translation step gene expression can also be tuned by using translation signals with different regulatory mechanisms and transcriptional strengths. These are controlled by the binding of various cellular translation factors at or near the translation signal in order to attract the ribosome for translation initiation. Depending on the organism, the mechanisms used by the ribosome to identify the translation initiation signal can vary significantly, which also affects the efficiency and dynamic range of translation regulation. In eukaryotes, ribosomes typically bind the methylated guanosine cap at the 5′-end of the protein-coding mRNA and then scan along the mRNA to identify the Kozak sequence and start codon for translation initiation <ref>. This scanning mechanism is highly influenced by the secondary structure in the 5′-UTR of the mRNA, especially around the start codon. It also explains why most eukaryotic mRNAs only encode a single protein coding sequence (monocistronic). Some eukaryotic viruses have mRNAs with several protein coding genes by bypassing the need for the 5′-cap binding and using instead an internal ribosome entry site (IRES) that mimics the 5′-cap, but can be present at internal sites within the mRNA. In contrast, ribosomes in prokaryotes bind the mRNA via RNA:RNA interactions that stabilize the ribosome onto the Shine-Dalgarno sequence within the ribosome binding site (RBS) in order to initiate the translation process. Given the strong dependence of translation initiation on RNA secondary structure, and several well-performing models available for its prediction, several tools have been developed to aid the bioengineer in *de*

novo design of translation signals for both eukaryotes and prokaryotes [31, 32]. Like transcription, translation can also be controlled by small-molecule riboswitches, although the repertoire of such RNA aptamer-binding small molecules is rather limited <ref>. Translation control is widely used by the bioengineer for optimization of multi-protein systems, often with severe cellular burden observed for high translation rates [33, 34].

Varying a combination of transcription-translation levels

Protein expression is a multi-step process, and can therefore be controlled at one or more of the steps involved. However, each step of biological information transfer has a different cost associated with it. According to one estimate replication, transcription, and translation are associated with 0.23%, 3.93%, and 95.83% ATP costs in the cell, respectively [35]. This is consistent with the observation that a cell has on average many more copies of proteins than mRNAs, and many more mRNAs than DNA [36]. However, in natural systems the relationships between copy numbers of DNA, mRNA, and proteins are often complex and specific to the protein in question [9, 37, 38, 4], reflecting the multi-objective nature of their optimisation over evolutionary time-scales. To gain additional insights into the expression burden of proteins while limiting the confounding regulatory and metabolic cross-talk with the chassis organism, several works have studied engineered systems that use a combination of transcription and translation signals to explore broad expression ranges [39, 40, 41].

Scott et al. established a phenomenological model to show how unnecessary protein expression reduces growth rate by reducing the available resources for the production of ribosomal / metabolic proteins [19]. In work by Ceroni and colleagues [39], the authors also found that overexpression of a protein by tuning transcription or translation results in reduced growth rate as well as reduced expression capacity, measured as the expression levels of a constitutively expressed "capacity monitor". Importantly, they showed that shortly after induction high-RBS combinations were less efficient than low-RBS combinations at producing the target protein. Using a ribosomal allocation model, the authors found that for the same GFP expression levels, high transcription (strong promoter) and low translation (weak RBS) results in higher efficiency of expression than the converse combination (weak promoter and high RBS).

However, a later study by Frumkin et al [40] found contradictory results. The authors constructed a library of ~14000 variants of GFP expressing constructs with a combination of constitutive promoters, RBSes, and GFP N-terminal sequences and measured the protein expression and growth rates of the library over 12 days. By analysing variants expressing the same amount of GFP, they identified those that grew significantly better than expected had higher "translation efficiency" (more proteins per mRNA molecule), driven by a stronger RBS. This was confirmed by Cambray and colleagues [41], who in an even larger study constructed and characterised a library of 244000 construct variants using a design-of-experiments approach to explore the effect of 8 sequence features on protein expression. Their results showed that for the most efficient trade-off between growth and protein production, the translation initiation rate must be balanced with its translation elongation rate to achieve high translation efficiency. They found that for similar protein expression levels, the lowest growth rate occurred in variants with the highest steady-state mRNA levels, indicating non-productive ribosome sequestration in "underachiever" constructs due to poor initiation on several copies of the mRNAs.

Solutions to problems

Problem 11.1 (Cost-benefit model for a metabolic enzyme)

The derivatives of cost and benefit function (i.e. the marginal costs and benefits, as functions of e) read

$$\begin{aligned} \frac{\mathrm{d}B}{\mathrm{d}e} &= \frac{v_{\max}\, a}{(e+a)^2} \\ \frac{\mathrm{d}H}{\mathrm{d}e} &= \gamma. \end{aligned} \tag{11.15}$$

For an extremal point, the two values must be equal. This yields

$$\begin{aligned} & \gamma\,(e+a)^2 = v_{\max}\,a \\ \Rightarrow \qquad & e = \pm\sqrt{\frac{v_{\max}\,a}{\gamma}} - a = \left(\pm\sqrt{\frac{v_{\max}}{a\,\gamma}} - 1\right)\,a. \end{aligned} \tag{11.16}$$

To obtain a positive solution for e, we discard the negative solution and require that the expression in brackets must be larger than 1, hence $v_{\max}/a > \gamma$. This means that in the point $e = 0$, the slope of $B(e)$ (given by $v_{\max}/a$) must be larger than the slope of $H(e)$ (given by γ). To prove that the extremum is an optimum, we need to show that the curvature of $F(e)$ is negative. Since $H(e)$ is linear, the fitness curvature is given by the second derivative of $B(e)$

$$\frac{\mathrm{d}^2 v}{\mathrm{d}e^2} = v_{\max}\,\frac{-a\,2\,(e+a)}{(e+a)^4} = -\frac{2\,v_{\max}\,a}{(e+a)^3}, \tag{11.17}$$

which is negative for all positive e. This shows that $F = v - H$ is negatively curved. In the case that $\frac{\partial(v-H)}{\partial e}$ is negative at $e = 0$ (and negatively curved everywhere), it will have a negative slope at all points $e > 0$, so $e = 0$ must be the maximum in this case.

Bibliography

[1] Frank J. Bruggeman, Maaike Remeijer, Maarten Droste, Luis Salinas, Meike Wortel, Robert Planqué, Herbert M. Sauro, Bas Teusink, and Hans V. Westerhoff. Whole-cell metabolic control analysis. *BioSystems*, 234:105067, 2023. doi: 10.1016/j.biosystems.2023.105067.

[2] Iraes Rabbers and Frank J. Bruggeman. Escherichia coli robustly expresses ATP synthase at growth rate-maximizing concentrations. *The FEBS Journal*, 289:4925–4934, 2022. doi: 10.1111/febs.16401.

[3] Mary J Dunlop, Zain Y Dossani, Heather L Szmidt, Hou Cheng Chu, Taek Soon Lee, Jay D Keasling, Masood Z Hadi, and Aindrila Mukhopadhyay. Engineering microbial biofuel tolerance and export using efflux pumps. *Molecular Systems Biology*, 7(1):487, 2011. doi: 10.1038/msb.2011.21.

[4] Gene-Wei Li, David Burkhardt, Carol Gross, and Jonathan S. Weissman. Quantifying absolute protein synthesis rates reveals principles underlying allocation of cellular resources. *Cell*, 157(3):624–635, 2014. doi: 10.1016/j.cell.2014.02.033.

[5] Leeat Keren, Jean Hausser, Maya Lotan-Pompan, Ilya Vainberg Slutskin, Hadas Alisar, Sivan Kaminski, Adina Weinberger, Uri Alon, Ron Milo, and Eran Segal. Massively parallel interrogation of the effects of gene expression levels on fitness. *Cell*, 166(5):1282–1294.e18, 2016. doi: 10.1016/j.cell.2016.07.024.

[6] Yuichi Eguchi, Koji Makanae, Tomohisa Hasunuma, Yuko Ishibashi, Keiji Kito, and Hisao Moriya. Estimating the protein burden limit of yeast cells by measuring the expression limits of glycolytic proteins. *eLife*, 2018. doi: 10.7554/eLife.34595.

[7] Wolfram Liebermeister, Elad Noor, Avi Flamholz, Dan Davidi, Jörg Bernhardt, and Ron Milo. Visual account of protein investment in cellular functions. *Proceedings of the National Academy of Sciences*, 111(23):8488–8493, June 2014. ISSN 1091-6490. doi: 10.1073/pnas.1314810111.

[8] William J Turner and Mary J Dunlop. Trade-offs in improving biofuel tolerance using combinations of efflux pumps. *ACS synthetic biology*, 4(10):1056–1063, 2015. doi: 10.1021/sb500307w.

[9] Yansheng Liu, Andreas Beyer, and Ruedi Aebersold. On the dependency of cellular protein levels on mRNA abundance. *Cell*, 165(3):535–550, 2016. doi: 10.1016/j.cell.2016.03.014.

[10] Rohan Balakrishnan, Matteo Mori, Igor Segota, Zhongge Zhang, Ruedi Aebersold, Christina Ludwig, and Terence Hwa. Principles of gene regulation quantitatively connect DNA to RNA and proteins in bacteria. *Science*, 378 (6624):eabk2066, 2022. doi: 10.1126/science.abk2066.

[11] Lukas Marschall, Patrick Sagmeister, and Christoph Herwig. Tunable recombinant protein expression in E. coli: promoter systems and genetic constraints. *Applied Microbiology and Biotechnology*, 101(2):501–512, 2016. doi: 10.1007/s00253-016-8045-z.

[12] Erez Dekel and Uri Alon. Optimality and evolutionary tuning of the expression level of a protein. *Nature*, 436 (7050):588–592, July 2005.

[13] Jens G. Reich. Zur Ökonomie im Proteinhaushalt der lebenden Zelle. *Biomed. Biochim. Acta*, 42(7/8):839–848, 1983.

[14] I. Shachrai, A. Zaslaver, U. Alon, and E. Dekel. Cost of unneeded proteins in E. coli is reduced after several generations in exponential growth. *Molecular Cell*, 38:1–10, 2010. doi: 10.1016/j.molcel.2010.04.015.

[15] Hisao Moriya, Yuki Shimizu-Yoshida, and Hiroaki Kitano. In vivo robustness analysis of cell division cycle genes in Saccharomyces cerevisiae. *PLoS genetics*, 2(7):e111, 2006. doi: 10.1371/journal.pgen.0020111.

[16] Matteo Mori, Terence Hwa, Olivier C. Martin, Andrea De Martino, and Enzo Marinari. Constrained allocation flux balance analysis. *PLoS computational biology*, 12(6):e1004913, 2016. doi: 10.1371/journal.pcbi.1004913.

[17] Anne Goelzer, Jan Muntel, Victor Chubukov, Matthieu Jules, Eric Prestel, Rolf Nölker, Mahendra Mariadassou, Stéphane Aymerich, Michael Hecker, Philippe Noirot, et al. Quantitative prediction of genome-wide resource allocation in bacteria. *Metabolic engineering*, 32:232–243, 2015. doi: 10.1016/j.ymben.2015.10.003.

[18] Oliver Bodeit, Inès Ben Samir, Jonathan R. Karr, Anne Goelzer, and Wolfram Liebermeister. Rbatools: a programming interface for resource balance analysis models. *Bioinformatics Advances*, (vbad056), 2023.

[19] Matthew Scott, Carl W Gunderson, Eduard M Mateescu, Zhongge Zhang, and Terence Hwa. Interdependence of cell growth and gene expression: origins and consequences. *Science*, 330(6007):1099–1102, 2010. doi: 10.1126/science.1192588.

[20] Griffin Chure and Jonas Cremer. An optimal regulation of fluxes dictates microbial growth in and out of steady state. *eLife*, 2023. doi: 10.7554/eLife.84878.

[21] Benjamin D. Towbin, Yael Korem, Anat Bren, Shany Doron, Rotem Sorek, and Uri Alon. Optimality and suboptimality in a bacterial growth law. *Nature Communications*, 8:article number 14123, 2017. doi: 10.1038/ncomms14123.

[22] Kirill Sechkar, Harrison Steel, Giansimone Perrino, and Guy-Bart Stan. A coarse-grained bacterial cell model for resource-aware analysis and design of synthetic gene circuits. *Nat. Commun.*, 15(1981):1–17, 2024. doi: 10.1038/s41467-024-46410-9.

[23] Chen Liao, Andrew E. Blanchard, and Ting Lu. An integrative circuit–host modelling framework for predicting synthetic gene network behaviours. *Nat. Microbiol.*, 2:1658–1666, 2017. doi: 10.1038/s41564-017-0022-5.

[24] François Bertaux, Jakob Ruess, and Grégory Batt. External control of microbial populations for bioproduction: A modeling and optimization viewpoint. *Current Opinion in Systems Biology*, 28:100394, 2021. doi: 10.1016/j.coisb.2021.100394.

[25] Moshe Kafri, Eyal Metzl-Raz, Ghil Jona, and Naama Barkai. The cost of protein production. *Cell Rep.*, 14(1):22–31, 2016. doi: 10.1016/j.celrep.2015.12.015.

[26] Manlu Zhu, Qian Wang, Haoyan Mu, Fei Han, Yanling Wang, and Xiongfeng Dai. A fitness trade-off between growth and survival governed by Spo0A-mediated proteome allocation constraints in Bacillus subtilis. *Sci. Adv.*, 9(39), 2023. doi: 10.1126/sciadv.adg9733.

[27] Nguyen Quoc Khanh Le, Edward Kien Yee Yapp, N. Nagasundaram, and Hui-Yuan Yeh. Classifying promoters by interpreting the hidden information of DNA sequences via deep learning and combination of continuous FastText N-grams. *Frontiers in Bioengineering and Biotechnology*, 7, 2019. doi: 10.3389/fbioe.2019.00305.

[28] Travis L. LaFleur, Ayaan Hossain, and Howard M. Salis. Automated model-predictive design of synthetic promoters to control transcriptional profiles in bacteria. *Nature Communications*, 13(1), 2022. doi: 10.1038/s41467-022-32829-5.

[29] Heidi Redden and Hal S. Alper. The development and characterization of synthetic minimal yeast promoters. *Nature Communications*, 6(1), 2015. doi: 10.1038/ncomms8810.

[30] Michael E. Lee, Anil Aswani, Audrey S. Han, Claire J. Tomlin, and John E. Dueber. Expression-level optimization of a multi-enzyme pathway in the absence of a high-throughput assay. *Nucleic Acids Research*, 41(22):10668–10678, 2013. doi: 10.1093/nar/gkt809.

[31] Tim Weenink, Jelle van der Hilst, Robert M McKiernan, and Tom Ellis. Design of RNA hairpin modules that predictably tune translation in yeast. *Synthetic Biology*, 3(1), 2018. doi: 10.1093/synbio/ysy019.

[32] Howard M Salis, Ethan A Mirsky, and Christopher A Voigt. Automated design of synthetic ribosome binding sites to control protein expression. *Nature Biotechnology*, 27(10):946–950, 2009. doi: 10.1038/nbt.1568.

[33] Lior Zelcbuch, Niv Antonovsky, Arren Bar-Even, Ayelet Levin-Karp, Uri Barenholz, Michal Dayagi, Wolfram Liebermeister, Avi Flamholz, Elad Noor, Shira Amram, Alexander Brandis, Tasneem Bareia, Ido Yofe, Halim Jubran, and Ron Milo. Spanning high-dimensional expression space using ribosome-binding site combinatorics. *Nucleic Acids Research*, 41(9):e98–e98, 2013. doi: 10.1093/nar/gkt151.

[34] Iman Farasat, Manish Kushwaha, Jason Collens, Michael Easterbrook, Matthew Guido, and Howard M Salis. Efficient search, mapping, and optimization of multiprotein genetic systems in diverse bacteria. *Molecular Systems Biology*, 10(6), 2014. doi: 10.15252/msb.20134955.

[35] Elisa Marquez-Zavala and Jose Utrilla. Engineering resource allocation in artificially minimized cells: Is genome reduction the best strategy? *Microbial Biotechnology*, 16(5):990–999, 2023. doi: 10.1111/1751-7915.14233.

[36] Victoria Munro, Van Kelly, Christoph B. Messner, and Georg Kustatscher. Cellular control of protein levels: A systems biology perspective. *PROTEOMICS*, 24(12–13), 2023. doi: 10.1002/pmic.202200220.

[37] Fredrik Edfors, Frida Danielsson, Björn M Hallström, Lukas Käll, Emma Lundberg, Fredrik Pontén, Björn Forsström, and Mathias Uhlén. Gene-specific correlation of RNA and protein levels in human cells and tissues. *Molecular Systems Biology*, 12(10), 2016. doi: 10.15252/msb.20167144.

[38] Yun-Chi Tang and Angelika Amon. Gene copy-number alterations: A cost-benefit analysis. *Cell*, 152(3):394–405, 2013. doi: 10.1016/j.cell.2012.11.043.

[39] Francesca Ceroni, Rhys Algar, Guy-Bart Stan, and Tom Ellis. Quantifying cellular capacity identifies gene expression designs with reduced burden. *Nature Methods*, 12(5):415–418, 2015. doi: 10.1038/nmeth.3339.

[40] Idan Frumkin, Dvir Schirman, Aviv Rotman, Fangfei Li, Liron Zahavi, Ernest Mordret, Omer Asraf, Song Wu, Sasha F. Levy, and Yitzhak Pilpel. Gene architectures that minimize cost of gene expression. *Molecular Cell*, 65(1):142–153, 2017. doi: 10.1016/j.molcel.2016.11.007.

[41] Guillaume Cambray, Joao C Guimaraes, and Adam Paul Arkin. Evaluation of 244, 000 synthetic sequences reveals design principles to optimize translation in Escherichia coli. *Nature Biotechnology*, 36(10):1005–1015, 2018. doi: 10.1038/nbt.4238.