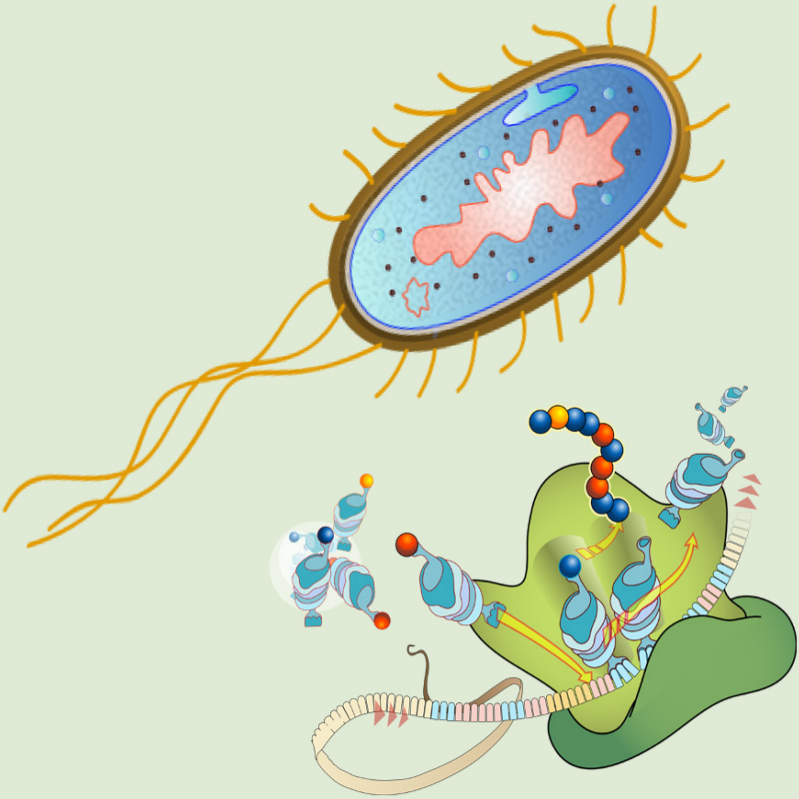
The Bitesize Bio Guide to Protein Expression



Victoria Doronina
Edited by Emily Crow

Contents

Chapter 1
Introduction

Your nightmare usually starts like this. Your boss tells you: "You have to purify protein X. Do a PCR, clone into vector Y, transform into *E. coli*, express and purify. I expect to have a gram of protein in two weeks".

Your PCR works a treat, and your cloning does too. You induce the cells with IPTG, run your samples on an SDS-PAGE gel and then stare blankly at multiple bands, desperately trying to identify a fat band that could be your overexpressed protein. Trying, and failing.

Next, you do a western blot and see multiple bands. Or, you see denatured insoluble material and discover that your protein is in inclusion bodies. Then and only then do you look at what might have gone wrong. With a list of possible fixes, you repeat the experiment over and over, making minor tweaks each time and eating up valuable months of project time.

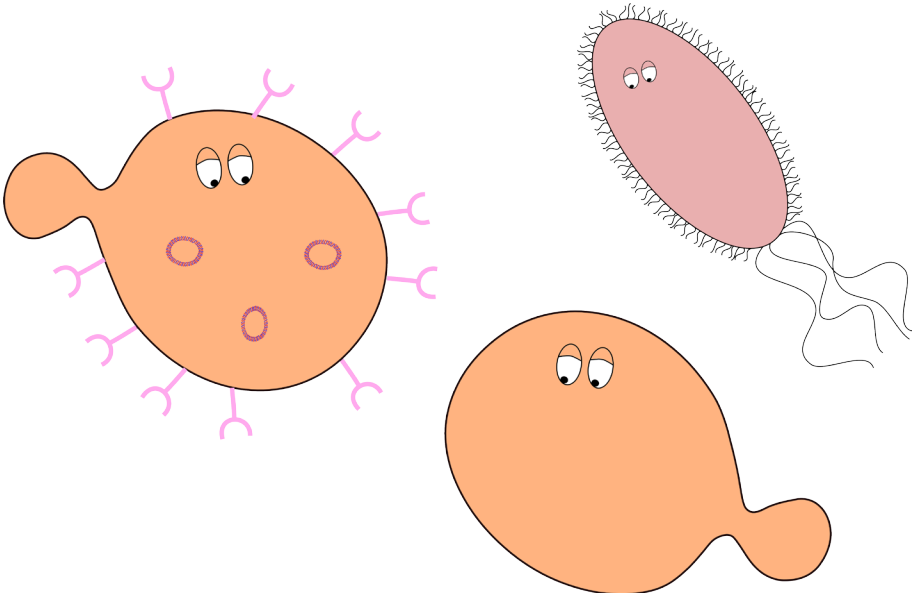
There is a joke about primatologists who decided to compare the problem-solving skills of a chimp and a biochemist. The chimp and the biochemist are put in separate rooms, each of which has a banana dangling from a rope (too high to reach) and some boxes in the corner. The chimp starts jumping but can't get the banana. It sits on a floor for a while thinking, then stands up, drags the boxes together, and piles them on top of each other until the pile is high enough to climb up and reach the banana. The biochemist, too, jumps, can't reach the banana, and sits down to reflect. Then he stands up and exclaims: "There no point in wasting time thinking, I must continue jumping!".

The moral of the story, of course, is that excessive "jumping" to obtain your purified protein can be prevented with some extra thought and careful planning. This book will deconstruct the

process and give you tips on how to design and troubleshoot protein expression systems. If you're just starting to plan your experiment, then the information provided here will help you spot potential problems and stop them before they happen. If you've already tried and failed to express your protein, then you can use this book as a guide to identify what might have gone wrong and how to fix it.

I am an experienced cellular biologist (Methuselah in postdoc years) with training in the uncool but extremely useful discipline of eukaryotic gene transcription and translation. I've seen what goes wrong with protein expression and can usually help solve the problem – troubleshooting is my hobby. Sometimes resolving your problems with protein expression are as simple as F.B.P. (follow the bloody protocol), but many times you will need to know how protein expression works to develop your own protocol. After reading this guide you should be able to put together a well-designed protein expression system and troubleshoot problems with your (and your colleagues') protein expression problems. This book will give you the tools you need to apply a simple algorithm – know how it works, find where it is broken, then fix it!

Chapter 2
Choosing an expression system



Following his recent transformation, John was really learning to express himself.

What is an expression system anyway?

There was a time not so long ago when genetic engineering was not just unheard of, but also undreamt of. But the pre-historic biochemists still managed to purify significant quantities of several proteins by sifting through tons of raw tissue in slaughterhouses. Medical doctor and writer Le Fanu, in his book "The Rise and Fall of Modern Medicine", even goes as far as to say that heterologous expression of the poster child for genetically engineered proteins – human insulin – was unnecessary, because there was a good source of it in porcine tissue. Thankfully, there's no need to go to this messy extreme anymore, as there are many techniques available for protein expression under laboratory conditions.

By "protein expression" I mean protein "overexpression", which is the synthesis of a protein in a greater quantity than would occur naturally from the native promoter. An (over)expression system consists of a master copy of your gene under the control of a high expression-level promoter and something that supplies the protein building blocks – this can be a live cell or a cell-free system. The simplest expression system is whatever cell your native protein comes from. It consists of two main elements: the gene (i.e. a set of instructions) and protein synthesis components nicely packaged inside a cell (i.e. the machinery to manufacture the protein). This native system will generate a relatively low-copy "artisan" product.

If you can grow your organism easily under laboratory conditions and your target protein is a high-abundance protein (such as translation factors and structural proteins), then you should strongly consider purifying it "as is" from the original organism. This will save you a lot of work cloning and dealing with

consequences of overexpression. No such luck? Let's talk about expression systems.

Choosing your heterologous expression system

You should choose both the vector and the expression system carefully, because the more difficult a protein is to express, the more control you will need over the timing and magnitude of expression. Please don't just start with whatever old plasmid and *E. coli* strain is knocking around your lab, as you may regret this very soon.

As a rule, you should keep as close to the native environment of your target protein as possible. In terms of trying different expression systems, you can go down the evolutionary ladder using a more simple system to produce a protein from a higher organism (for example, use *E. coli* to produce yeast proteins), but certainly not in the opposite direction. Think of it this way: even though you could produce a yeast protein in human cells, there would be no point, because human cells are harder to grow and the media are more expensive.

Is your protein from a prokaryote? If so, then *E. coli* is probably your best bet, as you will have a wide selection of plasmids and strains to choose from (*see A horse called Escherichia* for more details). Is your protein from a eukaryote? Then you will probably be better off with a eukaryotic expression system, ranging from yeast to human cells. These systems may be less familiar than *E. coli*, but have the advantage of producing a protein that is closer to the native state. And if you can work with *E. coli*, you can work with any cell culture system, honest.

Both prokaryotic and eukaryotic expression systems can be replaced with a corresponding cell-free system, which will allow you to bypass the *in vivo* expression constraints. Live cells constantly select proteins that "make sense" and inactivate defective proteins. If your artificial construct rings alarm bells in the cell indicating that it is an "incorrect product", it will be discarded. That is why it is often difficult to express a heterologous protein in sufficient quantities, and why they are often sequestered in inclusion bodies. A cell-free system allows you more flexibility in expressing proteins, while still generating a sufficient yield for most applications.

Vectors

The expression vector you will use is typically a plasmid, where your gene of interest is expressed from a strong (often viral) promoter – typically the T7 or SP6 bacteriophage promoter in *E. coli*, or the cytomegalovirus (CMV) promoter in mammalian cells. Strongest doesn't always mean best, and your protein could be toxic even under uninduced conditions, so do try to use a tightly controlled promoter. The vector can be low or high copy number. Again, more doesn't necessarily mean better: replicating and maintaining a plasmid places a burden on the cell.

Important elements of the vectors include selection markers (most frequently antibiotic resistance genes) and "maintenance sequences" – origin of replication, segregation and copy number-controlling elements. Vectors for expression in higher eukaryotes often contain other sequences necessary for successful expression (enhancers, etc.). It's a bad idea to mess with them.

Of course there are also vectors that allow you to integrate your construct into chromosomal DNA. On the plus side, this makes the

construct stable and may eliminate the need for constant selection. On the minus side (there is always a minus side), in eukaryotic cells, your protein production will depend on where your vector integrates, as heterochromatin will completely silence it.

In this chapter, we'll discuss the various expression systems that are available to you, what their advantages and disadvantages are and how to choose the right one for expressing your protein.

A horse called Escherichia

I think "Escherichia" is a lovely name for a pony, don't you agree? But *Escherichia coli* is not a pony but a molecular biology and biochemistry workhorse. We clone and amplify plasmids in *E. coli*, and it is very often the tool for protein expression. *E. coli* was one of the first model organisms, so we know a lot about protein expression and purification in this bacterium. It is also the simplest organism to grow in the lab. But "simple" doesn't always mean "right". As more and more proteins were expressed in *E. coli*, it became clear that it is not suitable for expression of many eukaryotic proteins. Even if you correct for different codon usage, *E. coli* lacks the machinery necessary for eukaryotic fine-tuning such as targeting to different compartments and various post-translational modifications – disulfide bonds formation, acetylation, glycosylation, etc.

If you're expressing a prokaryotic protein, though, *E. coli* is an excellent choice, and is actually a lot more flexible than many scientists are aware. It's a pity that the typical biochemist rarely knows more about *E. coli* than basic stuff like "grow at 37°C in LB". This is fine if you don't have any problems expressing your protein, but if you do, you are left with an insolvable (pun intended) puzzle. Knowing just a little more about your *E. coli* workhorse will help you express your heterologous protein.

There is more to E. coli *than BL21: Strain choice*

BL21 (DE3) is the standard *E. coli* strain used for protein expression. It supports expression from the strong T7 promoter, which is found in most classical expression vectors. But transcription from this promoter is "leaky", i.e. there is some level of expression even in the absence of induction. If the protein is toxic, this leaking faucet creates selective pressure to get rid of the protein, leading to the accumulation of mutations in the promoter or the cloned gene.

Protein aggregation is common in BL21, so it is worth trying other specialized strains that have tighter T7 promoter regulation. These strains are engineered to repress the basal level of T7 promoter expression by one of two methods: 1) controlling expression of the T7 polymerase as part of a tightly regulated operon (strain BL21-Al), or 2) introducing the pLysS plasmid, which carries the gene for T7 lysozyme, a natural inhibitor of T7 RNA polymerase (strain BL21 (DE3) pLysS).

There are also specialized, commercially available strains that are specifically designed to deal with other protein expression problems, for example:

- "BL21 plus" and "Rosetta", which correct for codon bias;
- "ArcticExpress", which improves protein folding due to chaperone overexpression; and
- "SHuffle" and "Origami", which promote disulfide bond formation for optimal folding.

Do try to shop around for the best strain.

Other parameters that you have at your disposal

E. coli grows in "rich" media containing cell extracts (most commonly buffered LB) or completely synthetic media (like M9) that contain a buffer and a carbon source. Depending on the application, you can choose to use a variety of different media. I strongly recommend checking out this protocol for generating high yields of recombinant proteins in *E. coli*.

Optimal *E. coli* growth occurs at 37°C, but it can actually grow at a range of temperatures from 14°C up to 49°C. Growing *E. coli* at temperatures higher than 37°C will lead to heat shock, which is not optimal for protein production, as the cells will spend most of their energy on heat-shock defense. However, you can fine-tune your *E. coli* growth and protein expression by lowering the temperature and slowing down expression, as reduced protein expression lowers the chances of protein aggregation. For example, the ArcticExpress strain mentioned above is typically grown at 14°C.

Fifty shades of yeast

If your gene of interest is eukaryotic and you want the protein to be functional, you may want to consider moving from *E. coli* to yeast. Yeast are as easy to work with as prokaryotic microorganisms, but with a difference – they are eukaryotes. This means that yeast can carry out all of the post-translational modifications that would occur in higher organisms, such as transport to different cellular compartments, precursor protein processing and glycosylation. This is important for expressing proteins that are native to eukaryotic organisms, as these modifications can be critical for protein solubility and activity.

If you can grow *E. coli*, then you will have no problem growing yeast. The growth conditions and other techniques are actually pretty similar:

- yeast can grow on agar plates (where they form colonies) or in liquid culture;
- they can grow in minimal media that contain little more than a carbon source (such as glucose) and a nitrogen source, or in rich media similar to LB;
- they can be transformed or electroporated with vectors that were cloned and amplified in *E. coli*.

...plus, yeast smell so much better than *E. coli*.

Granted, yeast grow noticeably slower than *E. coli* (they divide roughly every three hours instead of every thirty minutes). The other disadvantage is that you cannot just boil the cells to extract your protein. Yeast cells are surrounded by tough chitin-based cell walls that need to be broken with glass beads. You can break open large volumes of cells using familiar techniques like sonication or using a French press.

The species

There are several yeast species that are commonly used for protein purification. You've probably heard of *Saccharomyces cerevisiae* or baking yeast – you know bread, wine, beer? It has also been helping out in molecular biology labs for several decades now, so there are a lot of genetic and biochemical tools available for working with this strain. *Kluyveromyces lactis* and *Kluyveromyces marxianus* (*Figure 1*) are very similar to *S. cerevisiae*. These strains are used for industrial-scale production of some proteins. Vectors

FIGURE 1: *Kluyveromyces marxianus* colonies. Source: Ude. Licensed under CC BY-SA 3.0

for this species are commercially available. Finally, several *Pichia* species can grow on methanol as a cheap carbon source, and are used commercially and in the lab.

The only proteins I would not recommend expressing in yeast are secreted proteins, because yeast cells don't typically secrete many proteins. For secreted eukaryotic proteins, you'll want to use an expression system based on a higher organism, such as those described in the next few sections.

A very friendly armyworm

Now we are jumping from the cozy, simple world of microorganisms into cell culture. Thinking about protein expression in mammalian cells? Think again – insect cells may be better. Insect cells, baculovirus vectors and media are commercially available, and some universities even have insect-cell protein-expression facilities.

Cells derived from bugs are more "exotic" than yeast, but they have an advantage: they are not as difficult to break open. Plus, proteins expressed in insect cells are often more soluble than proteins expressed in *E. coli* due to various post-translational modifications such as glycosylation (adding sugars to proteins). Glycosylated proteins can be difficult to purify using traditional methods, but a lot of important human proteins are glycosylated, including many enzymes, hormones and antibodies. Expressing glycosylated proteins without their sugar decorations often leads to aggregation. There are several other post-translational

modifications that your insect cells will also gladly make for you, such as disulfide bonds and acylation.

Insect cell lines are particularly useful for expressing secreted proteins and some mammalian membrane proteins (which are notoriously difficult to express and purify). This is because insect cells have all of the same cell compartments as mammalian cells, but are faster and more flexible to work with compared to establishing a stable expression in a mammalian system. Plus, insect cells are less likely to become contaminated during culture.

Insect cell lines

The first generation of insect cell lines used for protein expression, the sort of "BL21 of insect cells", included Sf9 and Sf21. These cell lines were derived from the fall armyworm, a drab-looking moth (*Figure 2*). The more advanced insect cell lines (for example, Hi-5, which was derived from another moth, the cabbage looper) can grow without the increased carbon dioxide concentration that is necessary for human cells. They can grow at room temperature and don't require the addition of serum. Insect cells can grow in suspension (almost like microorganisms) or as a semi-adherent layer, which greatly increases protein production per milliliter of media.

FIGURE 2: *Spodoptera Frugiperda* worm. Source: Canadian Biodiversity Information Facility. Licensed under public domain.

Vectors for expressing heterologous proteins in insect cells

The vectors used for protein expression in insect cell lines are based on insect viruses and are not dangerous for humans or animals – this is yet another advantage of using an insect-cell expression system, from a biosafety point of view. These vectors are a hybrid between a double-stranded insect virus and a standard *E. coli* plasmid, so they can be cloned as usual in bacterial cells and amplified in insect cells. The gene of interest is typically cloned under a strong promoter, pPolh, which drives expression during the late stage of insect cell infection, producing a lot of protein – up to an unbelievable 1 g/mL of media. The vectors contain various standard purification tags (His, FLAG), so after the expression, it's purification as usual.

You can also construct a stable line, where the gene is inserted in the genome, just like with mammalian cells.

Popular (cell) culture

The last "live system" we are going to discuss is mammalian cells. Using mammalian cells to express mammalian proteins, an admittedly straightforward approach, is on the opposite side of the spectrum from the principle: "What is true for bacterial cell is true for the elephant" (ten extra points if you know who said this*). By using mammalian cells to express mammalian proteins, you jump right over potential problems with codon usage and defects in protein function due to differences in folding and other post-translational modifications.

On the minus side, media for growing mammalian cells are more expensive than media that support insect cell growth

* Jacques Monod. One hundred extra points if you know who Jacques Monod was.

because the purity must be much higher. In addition, mammalian cells are slow to divide compared to yeast, which makes protein expression a lengthier process. Finally, your aseptic technique needs to be impeccable when handling mammalian cells, otherwise the biologist's old friends (bacteria and yeast) will become your enemies.

Cell lines

When speaking about mammalian cell expression systems, I deliberately use the term "mammalian", not "human". While human cell lines such as HeLa (*Figure 3*) and HEK293 may be the most well-known cultured cells, they are not the best to work with. They are "immortal" cancer cells that can divide indefinitely, instead of for a fixed number of divisions, like normal cells. This is the result of mutations in genes that regulate the cell cycle, which leads to genome instability. Thus, they tend to lose genetic material during cultivation, so your painstakingly optimized protein expression levels can deteriorate over time.

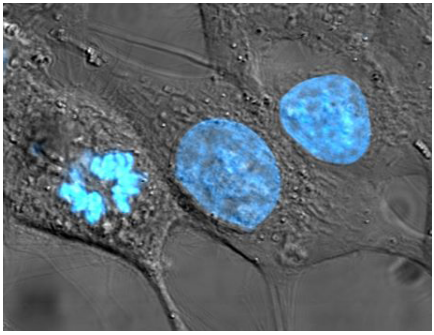


FIGURE 3: HeLa cells stained with Hoechst 33258. Source: TenOfAllTrades. Licensed under public domain.

For mammalian protein expression, cellular biotechnologists prefer to use rodent cell lines. For example, Chinese hamster ovary (CHO) cells can grow in suspension, e.g. in the entire volume of your medium, not just on a surface (as HeLa and most mammalian cell lines do). The NS40 cell line, which was derived from mice, is another example of a cell line that is used in industry to produce various proteins such as antibodies.

Vectors

The genetic engineering side of expressing proteins in mammalian cells is familiar – all you need is a vector, which can be amplified in *E. coli*, purified with extra care and then transformed into the cells. Expression of your gene of interest will typically be driven by a strong promoter from a mammalian virus such as SV40.

You have a choice between maintaining your construct in the cytoplasm or integrating it into the genome. The disadvantage of "free" vector is that you will need to keep your cells under permanent selection. Integrating your gene into the genome will create a number of stable cell lines, but the expression of your protein will vary hundreds and even thousand of times from clone to clone. This clonal difference in the level of protein synthesis depends on where your construct integrates: in areas of intense transcription, your gene will also be transcribed at high levels, but if it integrates into a rarely transcribed section of the genome, you will only see very low levels of expression. Do check a number of clones to make sure you find one with a high yield. The translation levels may also vary in a single clone depending on the cell culture growth stage, just as in microbes, so it's worth monitoring the levels of your protein during culture growth up to senescence.

After you've chosen a cell line and established expression levels, you can apply the same tricks as in any expression system to optimize expression, including varying different media components. But regardless of what vector and cell line you choose, be sure to freeze your cells early and often, and check the level of expression from time to time.

I want to be (cell) free

The difference between cell-based expression systems and cell-free systems (or cell extracts) is the same as the difference between fruit and a smoothie. Fruit is fruit, and a smoothie is smashed-up fruit. I prefer smoothies because I don't have to chew – but there's a place for fruit, too!

System overview

A cell-free system is a smashed-up cell extract that doesn't contain its own RNA but does contain all of the components necessary for protein synthesis. It can consist of either purified components that are necessary for transcription and translation (a reconstituted system) or simple crude extract. Reconstituted systems often allow you to select for your newly synthesized, untagged protein among tagged working components. Just fish out the tagged proteins involved in transcription and translation using an affinity resin, and you have a purified solution of your protein.

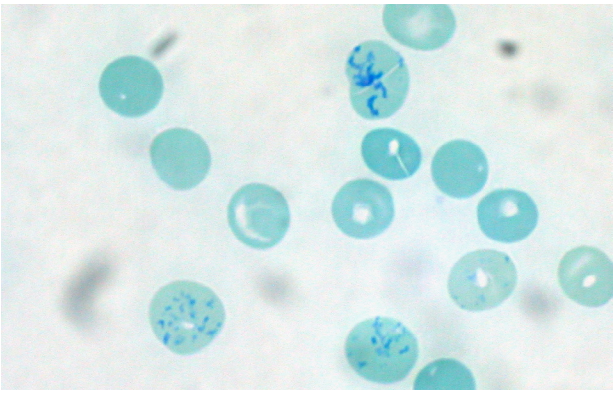


FIGURE 4: Human reticulocytes in a blood smear stained with supervital stain. Source: Ed Uthman, MD. Licensed under CC BY 3.0.

The more traditional, "crude" systems can be more robust than reconstituted systems – sort of like the difference between purebreds and mutts. Cell-free systems can be bacterial (usually made from *E. coli* cells)

or eukaryotic. The most commonly used eukaryotic systems are made from plant cells (i.e. wheat germ extract) or animal cells (i.e. rabbit reticulocyte lysate) (*Figure 4*, human reticulocytes). Yeast cell-free systems are generally not available, but unlike rabbit reticulocyte lysates, you can easily make them in the lab.

Regardless of the original cell type, cell-free systems come in two flavors:

- "linked systems", in which you synthesize and supply the mRNA, usually using a separate kit; and
- "coupled systems", in which supply only your expression vector – the system will both synthesize and translate the mRNA.

Coupled systems are more expensive, but the separate transcription step involved in using a linked system requires some RNA handling, which some people are (needlessly) afraid of.

Why would you use cell extracts?

If you need just a milligram of protein, there may be no point in optimizing an *in vivo* expression system: using cell-free extracts will be much quicker. There is no need to buy media, sterilize flasks or use the shaking incubators. Just defrost a tube, add your nucleic acid, put the mix on a hot block and get your protein several hours later.

On the other hand, if you did try *in vivo* expression and it failed epically, it is also worth trying *in vitro* system. Because there is no living cell to "mediate" between your gene and expression of your protein, there is no pressure on the system to circumvent protein

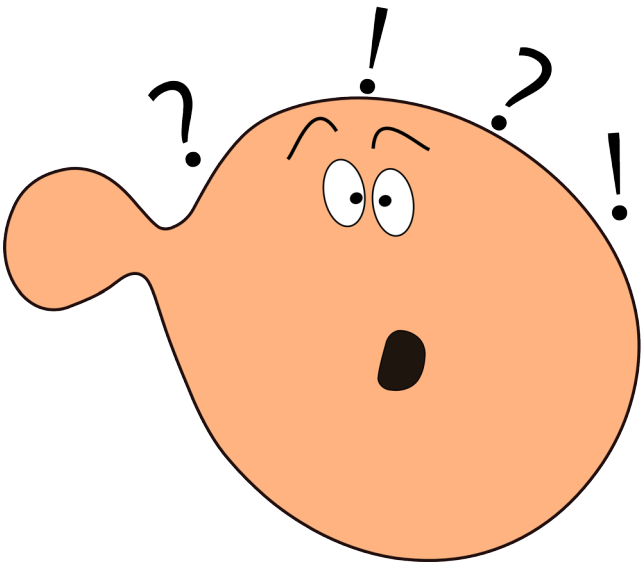
production by accumulating mutations in the gene or shuttling the protein into insoluble inclusion bodies. This is a significant advantage: you can express a toxic protein in a cell-free system. If your *in vivo* expression vector contains your gene of interest cloned downstream of a T7 or SP6 promoter, you can use it in an *in vitro* expression system, because the typical cell-free system contains the corresponding phage polymerases.

What if you express your protein in a cell-free system and the yield is not spectacular? It is possible to adjust some of the parameters in an *in vitro* system, such as the potassium and magnesium concentrations (for coupled systems) and temperature (all systems). You can also add in molecular chaperones or liposomes. Molecular chaperones assist in the folding of tricky proteins, and liposomes facilitate the expression of membrane proteins in their native form.

Another advantage of cell-free systems is the ability to incorporate "unnatural" amino acids (such as tagged or fluorescently labeled amino acids) into your protein, which will enable you to characterize it more quickly.

So with all of these advantages, is there any reason not to use a cell-free system? Well, the commercial cell-free extract systems are not cheap. However, they are definitely worth the cost for many applications.

Chapter 3
Exogenous protein expression: the problems



Ok, so by this point you've selected the very best expression system for your protein. The next thing you want to think about is any problems you may face expressing a heterologous protein in a non-native environment. Undergraduate studies leave most biologists with a belief that everything in the cell is – as a result of billions of years of evolution – beautifully optimized. This is broadly true, but you have to remember that the system is optimized for a particular purpose: specifically, the number of protein molecules that are produced under normal conditions. None of that "overexpression" or "heterologous expression" nonsense. When you force your expression system to churn out a huge number of identical protein molecules as quickly as possible, you create a number of bottlenecks. This chapter will help you to identify potential bottlenecks and design your experiment carefully to get around them. If you already have a problem with expression, checking what might have went wrong and comparing the *in silico* result with wet data will help you to fix it.

Why you need to know basic facts about your protein

Imagine that you are a detective looking for a suspect in a crowd. You start by constructing a profile – gender, approximate age, distinctive features – and matching it to people you see. If you think that he is a career criminal, you would visit places where such people congregate, like seedy bars (I imagine, never having had an inclination to go there).

To separate an individual from a crowd you need to know his specific features, and the same goes for your protein: to select an expression system for your protein, you need to know how to find your protein of interest in a crowded place – the cell. The main defining features of any protein are its molecular weight (MW),

isoelectric point (pI) and its "usual haunt" – the cell compartment where the mature protein is typically found.

You need to know the molecular weight of your target protein in order to plan what type of gel to run your samples on after expression and where to look for your protein on the gel. After you have retrieved your favorite gene sequence and converted it to the protein sequence, there are a number of programs that will help you predict the MW. I recommend a simple calculator from the reputable all-things-proteomics site ExPASy.

The isoelectric point is the pH where overall charge of your protein is close to 0. This is important to know because proteins tend to precipitate at their pI. Therefore, throughout protein expression and purification, you will have to avoid putting your protein in buffers that are close to the pI.

Protein domain prediction, which will tell you where your protein is likely to be found, allows you to plan the expression strategy. Is it normally secreted? This is good news, because you will not need to think about the best method to break the cells open and fish out your protein, but bad news, because you will need to concentrate it from the cultural medium. Is it trans-membrane? You are in for some hard work – but you didn't become a scientist to have an easy life, right? You should also be aware of whether or not the protein contains a cleaved signal peptide, as the protein will then appear as a persistently smaller band on your gel.

Now that you know basic facts about your protein, the next step would be how you are going to express it – or how to sidestep a total roadblock in expression.

Coding 101: Codon optimization

The DNA sequence is the first bottleneck in the pathway to protein overexpression. Even if you overexpress your gene in the original organism, not all codons for the same amino acid are the identical. Some of them encoded by rare tRNAs, which are not abundant enough to sustain protein overproduction. For example, in *E. coli* O157:H, leucine is encoded by six codons (*Table 1*).

TABLE 1: The frequency of codon usage for leucine in *E. coli*. Source: OpenWetware.org.

Codon	Frequency per thousand nucleotides
CUU	11.04
CUC	11.04
CUA	3.85
CUG	52.82
UUG	13.72
UUA	13.89

When the ribosome encounters these rare codons, it pauses... waits for the cognate tRNA... and can fall off, resulting in an incomplete protein, which will be degraded or accumulate in inclusion bodies. In this case, you will see normal transcription but little or no protein expression. To remedy this situation, you need to make sure that you have sufficient tRNAs to sustain protein production by changing rare codons to their more abundant cousins. This Codon Usage Database lists codon frequencies for large number of organisms. You can calculate codon usage in your gene of Interest using this program.

Problems with the DNA sequence can be corrected either by site-directed mutagenesis or by complete gene synthesis. Gene synthesis is not that expensive these days and usually includes sequence optimization.

Are you content with your gene content?

In addition to rare codons, another characteristic of your gene is critical for gene expression is its guanine-cytosine content (GC content). It varies even from gene to gene within a single organism, so it is always worth checking. The GC content of your gene is even more important if you express it in a different organism. If this is the case, you may have problems with cloning. AT-rich sequences are difficult to amplify, because they form only two bonds per nucleotide with the primer, while G and C form three. GC-rich sequences may form strong secondary structures and may be degraded by some *E. coli* strains.

After cloning you also need to know what to change in your sequence to adapt it to your expression system. The difference in GC content may adversely influence mRNA expression because the RNA polymerases in your expression system prefer the native GC content. You can determine whether your gene of interest will express well in your expression system using the GenScript Rare Codon Analysis Tool.

The GC content of a gene is a more global feature than rare codons, so if there is a significant discrepancy between the GC content of your gene and the expression system, you will need to correct by either synthesizing your gene or using a different expression system. Use this GC profile calculator to determine the GC content of large sequences or this simpler one for shorter sequences.

Repeat sequences: Lego blocks of secondary structure

Another macrofeature of your gene is repeat sequences, which can form strong secondary structures in RNA and stall translation. Repeats occur frequently in mammalian DNA (*Figure 5*), but the classic *E. coli* strains that are commonly used for protein expression hate repeat sequences and excise them. This will result in a defective protein or no protein production at all, and sequencing will show mutations in the coding sequence. Before you try expressing your protein in a heterologous system, look for repeat elements in the DNA for example using this program.

If you are expressing a gene with repeat sequences in a bacterial system in which recombination has been disabled, your DNA will remain intact, as the *E. coli* ribosome will plow on through them. But in a eukaryotic expression system, secondary structures in the RNA may cause problems with translation. If these secondary structures are located in the 5′ untranslated region of the gene, the repeat sequences will reduce translation initiation.

You can use this tool to predict possible secondary structures in your mRNA. RNA nucleotides are more promiscuous in their pairing, so almost every sequence will form structures, but structures with a formation energy of less than 10 kJ in the 5′ region and less than 50 kJ within the protein sequence should not cause a problem.

Prosite is a great site that contains links to several programs that detect primary sequence motifs. I recommend not ticking the "exclude patterns with a high probability of occurrence" option, as this option will show you potential sites for post-translational

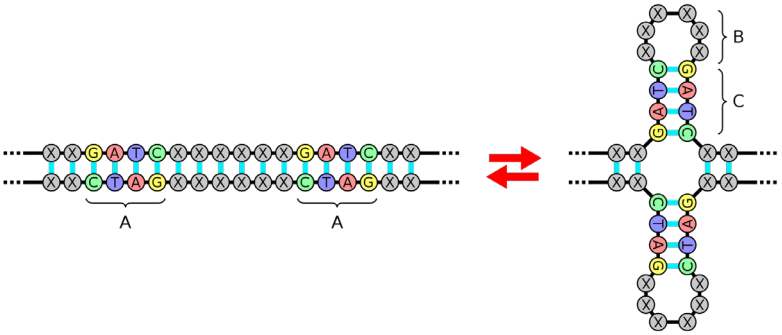


FIGURE 5: Palindromic DNA sequence, leading to the formation of secondary structure. Source: User: Acdx. Licensed under CC BY-SA 3.0.

modifications such as glycosylation and phosphorylation in your protein.

The game of introns

Introns are non-protein coding sequences in eukaryotic and some archaeal genes. They are sort of like conjunctions in the biological language – they can be eliminated without changing the meaning, and they're easy to get wrong. If you work with bacterial genes, you can express protein from the gene sequence as is, due to the absence of introns. However, if you are going to use a bacterial system, you will absolutely have to get rid of the introns in your sequence.

Working with cDNA will take care of potential splicing in eukaryotic systems. But do watch out for alternative splicing of eukaryotic genes, especially those involved in the immune system. Deleting predicted introns can activate secondary or alternative splicing sites leading to the expression of a protein isoform with different biological characteristics from the one you are studying. If you are using a eukaryotic expression system, remember to check for splicing sites and remove them. This website contains links to a

collection of tools for identifying splicing sites, including alternative splicing sites.

The importance of being soluble

Let's say you've designed a beautiful, streamlined construct to express your protein, and there are no problems with expression on the DNA or RNA level. Can you consider your ball already in the net/hoop/that square thing rugby ball goes into? Of course not: you've just finished the first level. The next level – or bottleneck, whichever you prefer – will be solubility of your protein.

For most purposes, "solubility" is a measure of how many molecules of your protein you can cram into an expression system so the protein stays in solution rather than crashing out of it, creating insoluble deposits (there are ways to deal with this, but that would be a different book).

Solubility depends on different factors such as the pI and MW of your protein, as well as the distribution of charged and hydrophobic amino acids. The more soluble your protein is, the more overexpression is possible, because insoluble molecules tend to clump together. During purification you can manipulate the solubility of your protein by changing the temperature, composition and pH of your buffer, but in the cell you have to put up with the intrinsic properties of the protein.

The solubility of a protein in a cell is not random – evolution has optimized the solubility of individual proteins depending on their function and abundance. In general, the more abundant your protein is under normal cellular conditions, the more soluble it will be. For example, a transcription factor with tens of copies per

cell will be much less soluble than some ribosomal proteins and histones that are common as dirt and just as highly expressed. The amino acid sequence of the transcription factor is not optimized for the high concentrations that result from overexpression, and the protein will likely become insoluble either in the cell or during purification.

There are a number of programs that will help you predict the solubility of your protein in different expression systems:

- The sequence-based PROtein SOlubility (PROSO) evaluator predicts which amino acids of your proteins are soluble during heterologous expression. Please note that this program doesn't work for transmembrane proteins.
- The MEMbrane protein EXperimentability (MEMEX) predictor predicts the solubility of membrane-spanning proteins. Bonus: it also calculates their "clonability" and "expressibility".
- The Recombinant Protein Solubility Prediction tool predicts the solubility of your protein when overexpressed in *E. coli*.
- Ticking 'SolPro' in the Scratch Protein predictor will return a prediction of your protein's solubility when overexpressed in *E. coli*.

If your insoluble protein ends up in inclusion bodies, all is not lost. These inclusion bodies are usually composed of 90% pure target protein, so there is still a chance of purifying your protein directly from the inclusion bodies. Check out the REFOLD database, which contains more than 1,100 protocols for refolding proteins.

The art of biological origami, or: Consider folding

Many proteins are folded by chaperones, and if translation proceeds too quickly in the heterologous system, this will result in incorrectly folded products, which manifest as inclusion bodies or inactive proteins. Even worse, some proteins are just unstructured and need other molecules to be present in order to adopt the correct structure. In the rare case that you need the unfolded polypeptide, rather than the folded state of your protein – for example, for antibody production – you may be better off with peptide synthesis. But in the majority of cases you have to have a properly folded, 3D protein.

Mess with protein sequence at your peril

Chopping off bits of a protein is a good way of completely messing up the protein structure. Inclusion bodies often form when unfolded (and therefore hydrophobic) polypeptide chains stick together. But in some cases, even a single amino acid substitution can lead to the same result – an unfolded protein.

An unfolded (or incorrectly folded) protein will result in familiar symptoms, including no protein as a result of degradation or the formation of protein aggregates. In this case you will be able to detect mRNA expression, but most of the protein will be in the insoluble fraction. If you manage to purify some protein, you will discover its improper behavior when you measure the physicochemical properties of your protein (for instance by circular dichroism).

To check that you are not dramatically affecting the structure of your target protein, you can use FoldIndex, which assesses the probability of folding for the provided sequence.

Not so fast: How speed of expression can screw up your protein folding

Several factors can lead to an overexpressed protein not folding properly. In real life, the protein is folded as it is translated. Some secondary structures form inside the ribosome during translation. The protein domains interact with each other as they emerge from the ribosome in the correct folding pathway sequence (*Figure 6*).

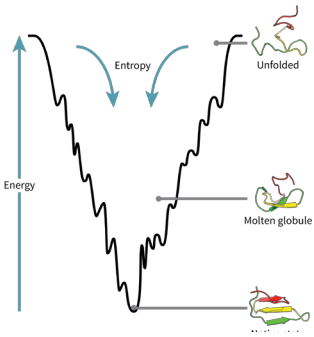


FIGURE 6: Protein folding energy funnel. The correctly folded form of a protein maximizes its free energy. Source: Thomas Splett-toesser. Licensed under CC BY-SA 3.0.

If translation is going too fast – for example, from a very strong phage promoter in an *E. coli* expression system – the protein might not have a chance to form the correct intermediates and will become trapped in an incorrect state, exposing hydrophobic domains. If your protein is forming aggregates, this could be the reason. In this case, you can try slowing the system down by changing the expression conditions to make them suboptimal – for example, lowering the temperature of the expression system. If this doesn't help, try changing the promoter to one that drives lower (slower) levels of protein expression.

But I couldn't possibly fold without my chaperone!

Some proteins need help folding, even under native conditions. This assistance is normally provided by protein folding helpers, which are known as chaperones. A shortage of chaperones will also lead to the accumulation of protein aggregates. If slowing down

translation didn't help, the next thing to try would be increasing the supply of chaperones.

Why won't my freaky protein fold?

Often, it is not the expression conditions which maketh the mess but the protein itself. The biochemistry textbooks tell us that all proteins have a secondary structure consisting of helices and sheets separated by loops. But there is a growing class of proteins known as intrinsically disordered proteins (IDPs) (*Figure 7*), which have an overrepresentation of polar and charged amino acids but lack aromatic and aliphatic residues. As a result, there is no hydrophobic core to serve as a nucleus of a stable tertiary structure under physiological conditions.

IDPs require intermolecular interactions for folding and can form different structures depending on the specific interaction. If you delete an unstructured domain and the protein becomes insoluble, you've been IDP'd. However, protein aggregation during IDP overexpression can be prevented by co-expression of a binding partner – for instance, a nucleic acid or protein.

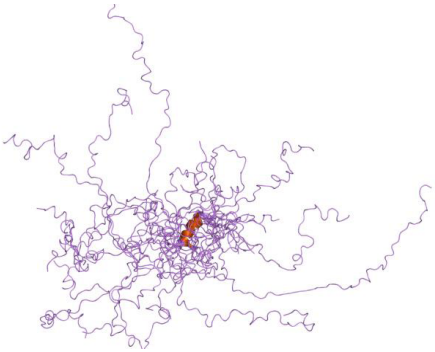


FIGURE 7: Cartoon representation of thylakoid soluble phosphoprotein Tsp9, an intrinsically disordered protein. Source: Jawahar Swaminathan and MSD staff at the European Bioinformatics Institute. Licensed under public domain.

You can check if your protein is intrinsically disordered using Disprot, a curated database of intrinsically disordered proteins. This site also has a list of methods for IDP prediction.

Try FoldUnfold to predict whether your peptide chain contains disordered regions.

Return to sender: Expressing a protein in the wrong compartment

Once, while travelling from Edinburgh to Aberdeen, I accidentally missed a coach change and ended up a hundred miles north of Aberdeen – in Inverness. At midnight. Surrounded by hungry wolves... Well, maybe there weren't any wolves, but this illustrates the point of this section: your problems with protein expression could be caused by incorrect localization.

While protein localization is crucial in eukaryotic cells, even bacterial proteins occupy specific locations within the cell. For example, cell division proteins are associated with the cell wall, and membrane proteins are definitely in a different compartment from cytoplasmic proteins. Gram-negative bacteria such as *E. coli* have an additional compartment between the inner and outer membranes known as the periplasmic space, where your (allegedly) secreted protein can end up if you mess with the protein folding (*see* The art of biological origami, or: Consider folding).

Proteins that are native to eukaryotic cells have more options than proteins expressed in bacterial cells. In a eukaryotic cell, proteins can be transported to, folded in and form disulfide bonds in the endoplasmic reticulum, mitochondria, chloroplasts or nucleus. They can also be inserted into the membrane or secreted. Tampering with secretion sequences, which are usually located in the N-terminal part of the protein, may prevent correct folding and result in very low levels of protein production as a result of protein degradation. When artificially overexpressing your protein, make sure that the tag you are going to use for protein affinity

purification – such as TAP or His6 – does not interfere with protein localization.

How do you determine where your protein is normally localized? As a cell biologist, I strongly believe that nothing beats direct observation of the protein in the cell, but that is a complete kettle of FISH. You may want to try using a prediction program instead:

- PSORT.org provides access to a number of programs that predict protein localization.
- TargetP specializes in eukaryotic proteins.

Export your protein

If your protein is secreted into the media, this will save you the trouble breaking down the cells and separating the protein from the thick soup of other proteins. This also makes affinity purification and concentration easier. In fact, many industrial companies prefer to work with secreted proteins. If you work with a secreted protein, make sure that you retain all the signal sequences. If you don't, you can use vectors, which allow to attach a secretion signal sequence to your protein of choice.

What if the expressed protein is toxic?

Despite how elegantly streamlined your code is, you may still experience problems when you move from in silico planning to "wet" conditions – i.e. actually expressing your protein. Nothing can prepare you for The Toxic Protein Problem. A protein that is toxic to the cells in your expression system is like chewing gum on

your shoe: you never know when you will encounter one, and it can be hard to get rid of.

The signs of a toxic protein are:

- Two types of colonies or cells after transformation: large and small. The large colonies or cells don't produce protein, and the small ones either cannot be grown or revert to the large phenotype (and produce no protein).
- The colonies or cells that do express your protein grow slowly or stop growing after induction of protein expression.
- The protein accumulates in insoluble aggregates.

If you are expressing a toxic protein from a leaky promoter, some protein will be produced even without induction, and this will be enough to make your cells sick. In this case, you will be able to detect transcription in uninduced cells, with little or no protein expressed from the sick cells.

To solve The Toxic Protein Problem, start by making the simplest changes to your expression system, gradually moving on to more serious changes if you're still struggling to express the protein:

1. First of all, change your expression conditions. Try lowering the temperature, or use a less rich medium (for example, change the carbon source). This will slow cell metabolism, including the speed of protein synthesis, lowering the amount of toxic protein produced and giving the cells time to adapt.
2. If this doesn't work, use a lower copy number vector or a weaker promoter: both of these strategies will reduce the

amount of toxic protein produced. Alternatively, you could use a more tightly controlled promoter. This will allow you to grow up your cells to a high density, then induce protein expression - which will kill the cells, but let you collect the protein (as viruses do). This is probably the most efficient way to deal with a toxic protein.

3. If you're working in *E. coli*, try using a different strain (*see* A horse called Escherichia).

4. As a last resort, you may need to change your expression system to something completely different. The most radical option here is to move to an *in vitro* system, where maintaining cell viability is not an issue (see I want to be (cell) free).

Conclusion

The take home message of this book is that a protein expression problem is complex – there is no single fix because it's not one problem, but a number of different problems, which require different solutions. With careful planning and preparation, you can achieve high expression levels of your protein, but remember that if things don't always go right the first time, and that there is always something else to try.

Farewell, reader. It's time for me to leave you to fend for yourself in the jungle of protein expression. I know you can succeed though. Just remember the monkey; don't keep jumping for the banana... take a step back, consider your position and your next options at each stage very carefully before diving in. That – along with knowing what your position and options are – is the real key to success in protein expression.

Please let me know about dead links in the book, what you didn't like, and what's missing. You can also share your joy if the book helped you in anything at all, even just to realize that books on molecular biology aren't always formal and full of impenetrable details. Please write to:

bsbproteinexpressionguide@gmail.com

Useful literature

Chapter 1

1. Brondyk WH. (2009) **Selecting an Appropriate Method for Expressing a Recombinant Protein**. *Methods Enzymol.* 463:131–47. DOI: 10.1016/S0076-6879(09)63011-1

Chapter 2

2. Carrió MM, Villaverde A. (2001) **Protein aggregation as bacterial inclusion bodies is reversible.** *FEBS letters* 489:29–33. PMID: 11231008
3. Chisti Y. (2003) **Metabolic engineering: debottlenecking metabolic networks.** *Biotechnol. Adv.* 21:295–6. DOI: 10.1016/S0734-9750(03)00039-9
4. Murray V, Huang Y, Chen J, Wang J, Li Q. (2012). **A novel bacterial expression method with optimized parameters for very high yield production of triple-labeled proteins.** *Methods Mol. Biol.* 831:1–18. PMID: 22167665
5. Sivashanmugam A, Murray V, Cui C, Zhang Y, Wang J, Li Q. (2009) **Practical protocols for production of very high yields of recombinant proteins using Escherichia coli.** *Protein Sci.* 18:936–48. PMID: 19384993
6. Fonseca GG, Heinzle E, Wittmann C, Gombert AK. **The yeast Kluyveromyces marxianus and its biotechnological potential.** *Appl. Microbiol. Biotechnol.* 79:339–54. DOI: 10.1007/s00253-008-1458-6.
7. Jarvis DL. (2014) **Recombinant protein expression in baculovirus-infected insect cells.** *Methods Enzymol.* 536:149–63. DOI: 10.1016/B978-0-12-420070-8.00013-1.
8. Gray D. (2001) **Overview of protein expression by mammalian cells.** *Curr. Protoc. Protein Sci.* 10:5.9.1–5.9.18. DOI: 10.1002/0471140864.ps0509s10.

Chapter 3

9. Gao F, Zhang CT. (2006) **GC-Profile: a web-based tool for visualizing and analyzing the variation of GC content in genomic sequences**. *Nucleic Acids Res.* 34:W686–91.
10. Dunker A, Silman I, Uversky V, Sussman J. (2008) F**unction and structure of inherently disordered proteins.** *Curr. Opin. Struct. Biol.* 18:756–764. DOI: 10.1016/j.sbi.2008.10.002

11. Uversky V. (2013) **Unusual biophysics of intrinsically disordered proteins.** *Biochim. Biophys. Acta.* 1834:932–951 DOI: 10.1016/j.bbapap.2012.12.008

About the author: Victoria Doronina

A product of Soviet no-nonsense science education, which culminated in a "red diploma" (University Degree in Microbiology), Vicki did her PhD in Molecular Biology at the University of Edinburgh. She has been working as a postdoc in several Russel group UK universities, while honing her skills in scientific and creative writing.

A Wikimedia Community Fellow (2011), she is an active participant in Wikimedia Movement and generally interested in Web 2.0