BIOL591: Introduction to Bioinformatics (2003)
Protein Overexpression: Structure vs. Function

I. Overview of protein overexpression

Take a look at that Pepsi (non-diet): high fructose corn syrup. Well, corn syrup doesn't have an appreciable amount of fructose in it, just lots of glucose. So where does the fructose come from? It comes from the enzyme glucose isomerase (originally from ground up *Bacillus coagulans*), which equilibrates glucose to glucose + fructose. Considering the gazillion liters of soft drink in the world, that's a lot of enzyme that's required.

Or what about human insulin, which since 1982 has been available for use by diabetics? Where does ***that*** come from? Not from ground up babies.

These and many other proteins are produced within bacteria for industrial use. Industrial use implies industrial quantities, implying a need to get bacteria to produce more of the desired protein than nature might otherwise dictate. The overproduction of protein raises several issues, some of which are:

- How can the gene of a eukaryote be expressed within a bacterium?
- How can the gene encoding the desired protein be transcribed at a very high rate?
- How can the transcript be translated at a very high rate?
- How can an unnaturally large amount of protein in small cells be convinced to maintain normal solubility?

After considering these questions, we'll turn to the question that will be of the greatest concern to us:

- How can bioinformatics contribute to the enhancement of protein production by microbes?

I.A. How can the gene of a eukaryote be expressed within a bacterium?

Antibodies, certain peptide hormones, and many other human protein are not naturally found in bacteria. If we wish to produce large quantities of them, there are a number of choices. One choice is not human tissue culture – it isn't possible to grow up enough economically. One can turn instead to recombinant cows (express the genes such that the protein appears in the cows' milk), plants, yeast, or bacteria. Since humans have had the greatest experience since antiquity growing microbes for industrial purposes (e.g. various fermentative processes), microbes are the most familiar choice. Bacteria pose a number of challenges in expressing eukaryotic genes:

- Eukaryotes and prokaryotes have different mechanisms for initiating transcription. A eukaryotic gene to be expressed in a prokaryote must be supplied with a prokaryotic promoter
- They also have different mechanisms for initiating translation. A eukaryotic gene to be expressed in a prokaryote must be supplied with a prokaryotic ribosome binding site.

- While the genetic code in prokaryotes is the same as in eukaryotes, some protein are modified after translation, and these modifications will not be found in eukaryotic proteins expressed in prokaryotes. In those cases where modification is important, bacteria are excluded as sites of production.

II.B. How to express large amounts of protein in bacteria

On one hand, you'd like to make as much protein as possible, so as to harvest the greatest amount of product in the least time. On the other hand, stuffing a bacterium with foreign protein is not compatible with good growth. To get rapid growth, you'd like to have little expression of the foreign protein. How to reconcile these two opposite needs?

Transcriptional fusions have proven useful in helping us regulate the amount of foreign protein expressed in a bacterial cell. One of the oldest yet still amongst the most effective ways of doing this is to fuse a souped up version of the regulatory region of the *lacZ* gene to the foreign gene to be expressed (Figure 1). Once this is accomplished, the foreign gene will not be expressed when the *lac* repressor binds to the regulatory region but will be highly expressed when the repressor is removed. Since the binding of the repressor is sensitive to sugars in the environment, the process is under our control.

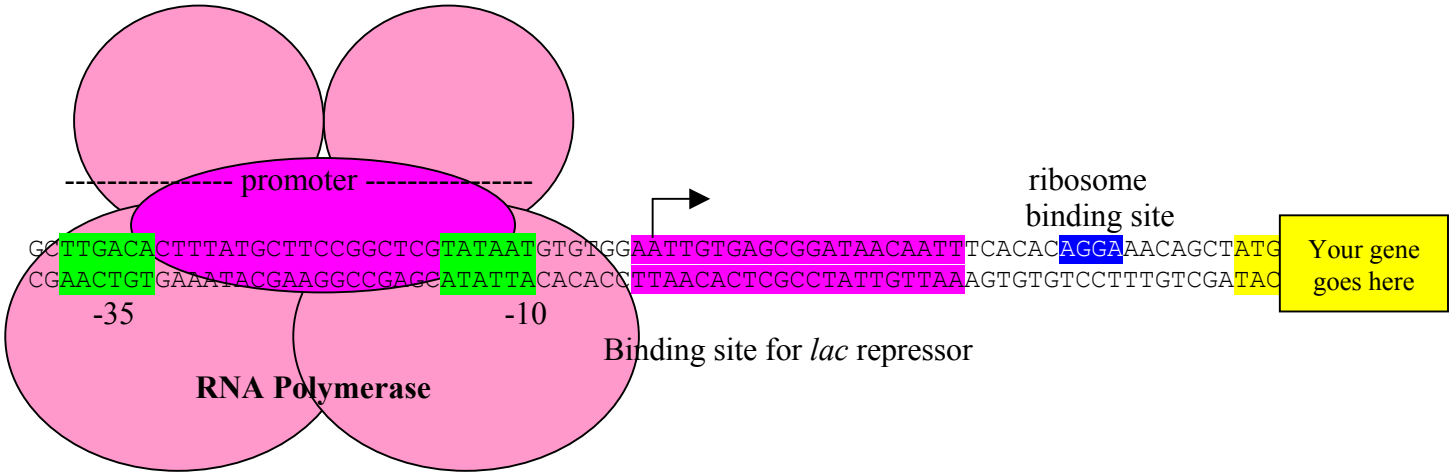

Fig. 1. Transcriptional Fusion between the *lac* regulatory region and a foreign protein.

SQ1. How can transcriptional fusion with the *lac* regulatory region help express foreign protein?

II.C. Problems with solubility

In general, if the protein is transcribed and translated, it will fold on its own to the three dimensional structure that gives it activity. However, expressing an abnormally high level of protein in a cell can lead to unwanted structures for two reasons. First, in some cases proteins need help in folding, from auxiliary proteins called chaperones. Very high expression of a protein can sometimes saturate the endogenous level of chaperones. Heightened expression of chaperones can help solve this problem. Second, weak interactions between protein that are insignificant at normal concentrations may become overwhelming when the cell is stuffed with the protein. This can lead to precipitation and formation of what are called inclusion bodies. In general, protein that have precipitated are not active. Techniques have been developed to recover activity from misfolded or precipitated protein, but these are tedious and not suitable for some automated processes.

III. The Scenario: Examine the structure of an overexpressed protein

III.A. The goal: Overproduced enzyme

As you read in the Scenario, you're on track to make an industrial amount of artificial cartilage. To do this, you'll need to make a major component of cartilage, chondroitin sulfate. It is glycosaminoglycan that is an important component of proteoglycans (Figure 2), complexes of proteins and sugars that provide structural support and compressive resistance in connective tissues such as cartilage. Chondroitin sulfate along with the other components of proteoglycans are used extensively in the field of tissue engineering to construct three-dimensional scaffolds that help direct and encourage the repair of damaged connective tissues, such as cartilage.

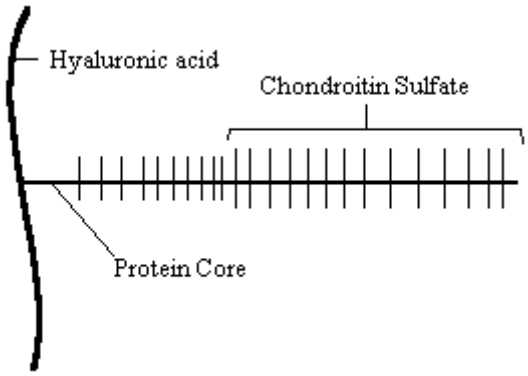

Figure 2: Proteoglycan Structure

One major component of chondroitin sulfate is glucuronic acid, added to the glycan in the form of UDP-glucuronate. To make lots of chondroitin sulfate, you'll need lots of UDP-glucuronate, preferably from a relatively cheap source. Organic synthesis of the compound is prohibitively expensive, so you turn to enzymes. The biological method of making UDP-glucuronate is to use the reaction illustrated in Figure 3, catalyzed by the enzyme UDP-glucose dehydrogenase. Thankfully, the substrate for the reaction is a derivative of glucose, dirt cheap (with the help of an enzyme that puts UDP on glucose).

With this in mind, you would like to produce a lot of UDP-glucose dehydrogenase, as part of the process of synthesizing chondroitin sulfate for use in tissue engineered scaffolds.

You isolate the gene encoding UDP-glucose dehydrogenase from the bacterium *Mesorhizobium loti* (chosen because of the superior qualities of the enzyme in industrial applications). The gene is transcriptionally fused to a modified *lac* regulatory region, as shown in Figure 1 and cloned in *E. coli.* DNA cloning involves the insertion of a gene into a self-replicating genetic element, typically a plasmid. A plasmid is an extra-chromosomal, circular piece of double stranded DNA found in bacteria. During cloning, a DNA fragment containing the gene of interest is joined to plasmid DNA by digestion with restriction enzymes. The vector DNA and the DNA fragment with the gene are cut by the same restriction enzyme, leaving complementary cohesive ends on each strand. The ends from each DNA fragment base pair, causing the DNA fragments to anneal and to form a recombinant DNA molecule. Next, this vector with the inserted gene is introduced

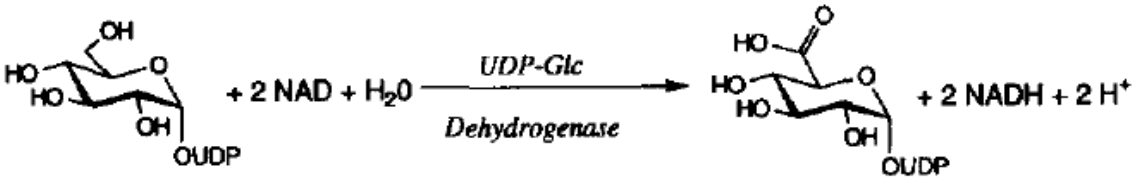

Figure 3: **Reaction Catalyzed by UDP-Glucose Dehydrogenase** (De Luca, C. et al. *Bioorganic & Medicinal Chemistry,* Vol. 4, No. 1, pp. 131-142, 1996).

into bacterial cells through transformation. The bacterial cells replicate the newly acquired DNA along with their own genome during repeated cellular divisions. Markers on the vector such as antibiotic resistance are used to ensure that the bacterium takes up and retains the plasmid. containing the gene of interest.

Now you have as much *E. coli* as you want expressing large amounts of UDP-glucose dehydrogenase.

SQ2. What is the function of UDP-glucose dehydrogenase?

SQ3. Can you visualize each step in the cloning process?

SQ4. What was the purpose of cloning the gene?

III.B. Overcoming precipitation through random mutagenesis

Unfortunately, while there is lots of the protein in the *E. coli* you've constructed, there's very little UDP-glucose dehydrogenase ***activity***. The protein is being made in an inactive form. A look through the microscope reveals crystalline bodies – inclusion bodies – that resolves the mystery. The overexpressed protein is precipitating, removing it from any possibility of catalyzing chemical reactions.

You reason thusly: if precipitation is caused by weak interactions between copies of the protein, perhaps you can prevent this by altering the amino acids on the enzyme's surface. Which amino acids to change? If you knew the answer to this question, you could perform site-specific mutagenesis of specific codons to alter the key amino acid residues. But you don't, so you try pot luck: random mutagenesis.

If you tried to make mutant enzyme by exposing the *E. coli* to a mutagen (like UV light or a chemical mutagen), you'd expose every gene in the cell to the same process. Either you'd use so little mutagen that mutations in all genes would be rare (and you'd have a difficult time finding the mutant enzyme you're after), or you'd increase the amount of mutagen and end up killing the cell. You need a method to mutagenize specifically the gene in question.

There are many ways of randomly mutagenizing a specific gene. One simple means is to use PCR to amplify the gene and then reinsert it into the plasmid. PCR employs a heat-stable DNA polymerase, often *Taq* polymerase (derived from the hot springs bacterium *Thermus aequalis*). Unlike the DNA polymerase used to routinely replicate DNA, *Taq* polymerase does not have proof-reading functions, hence it makes frequent mistakes. By frequent, I mean about one error for every 10^5 nucleotides replicated. This may not seem like much, but the DNA polymerase used in replication has an error rate of one mutation in 10^9 nucleotides!

SQ5. What is the expected number of mutations suffered by the gene encoding UDP-glucose dehydrogenase (roughly 1200 nucleotides) after amplification for 25 rounds of PCR? (Each round of PCR doubles the amount of DNA). What *fraction* of the genes do you expect to have at least one mutation? (You can answer these questions approximately or as precisely as you like)

Maybe one or more of the mutations will prevent precipitation, but how will you find out? You're ***not*** going to put 1000's of mutants under the microscope looking for inclusion bodies! (and the thought of doing 1000's of enzyme assays is no better). You need a surrogate method,

something that is easier to measure than the presence of inclusion bodies and is correlated to their presence.

Here's the trick. If any part of UDP-glucose dehydrogenase precipitates, it brings down the entire protein. If you attach to the enzyme another protein, one that is easy to measure, then it may be brought down also, losing activity. One of the easiest proteins to measure is green fluorescent protein (GFP). When the protein is active, it fluoresces green. When it isn't, then no fluorescence. If you make a ***translational fusion*** of the gene encoding UDP-glucose dehydrogenase to the gene encoding GFP, then the fate of both proteins are linked. You could easily screen 1000's of colonies on a single plate looking for heightened fluorescence, and those colonies showing high fluorescence should also have high UDP-glucose dehydrogenase activity. This would have to be verified experimentally, but you aren't averse to doing a ***few*** enzyme assays.

Using this trick, you find a couple of dozen mutant *E. coli* with high fluorescence. You find that it also has more UDP-glucose dehydrogenase activity, but not as much as you had hoped. Under the microscope, the inclusion bodies are still there in the mutants, though not as large as in the original strain.

So the trick helped, but not enough.

III.C. Overcoming precipitation through directed mutagenesis

Maybe it ***did*** help enough. You didn't get a lot of mutations, but maybe a mutation is possible that ***would*** prevent precipitation, just one not in your collection. Perhaps you can rationalize the mutations you found in terms of the positions of the altered amino acids on the protein's structure (sounds like one of the problems in the first problem set). If so, then you could predict what region of the protein is the place to mutate and use site-directed mutagenesis to specifically mutate those amino acids.

You therefore sequence the UDP-glucose dehydrogenase genes from the mutant *E. coli* and find which amino acids have been mutated. Unfortunately, they're spread out throughout the amino acid sequence. This bothers you for a while until you recall that the three dimensional structure of a protein may bring together amino acids that are distant from each other in the amino acid sequence. If you could see the three dimensional structure of this protein…

The three dimensional structure of UDP-glucose dehydrogenase from *Mesorhizobium loti* is not known, and it takes a lot of work and a lot of time to get a three dimensional structure of a protein. Is there an alternative?

What ***is*** available is the three dimensional structure of another UDP-glucose dehydrogenase, that from *Streptococcus pyogenes*. The amino acid sequences of these to proteins are sort of similar but by no means identical. Can this structure be of use to you?

This is the question we'll explore.

SQ6: What conditions call for random mutagenesis? Site-specific mutagenesis?

SQ7: How does error-prone PCR result in altered genes?

SQ8: What two types of gene fusions played important roles in this scenario? How do they differ?