10.492 - Integrated Chemical Engineering (ICE) Topics: Biocatalysis
MIT Chemical Engineering Department
Instructor: Professor Kristala Prather
Fall 2004

Lecture #9 – Biological Solutions to Process Design Problems

Handout: (1) Burton et al paper [*Nature Biotech*. (2002). **20**:37-45], (2) Biological Solutions grid, (3) Case Studies cont. handouts (three)

We previously discussed a series of process design challenges and defined them as either engineering or biological. We also discussed engineering solutions to those problems. Over the next few lectures, and to wrap-up the course, we'll talk about *biological* solutions to some of those problems and to the others that we have not yet discussed.

Problem/Solutions Grid

	Engineering Solution	**Biological Solution**
Engineering Problem	Yes	No (*1 case exception)
Biological Problem	Yes	Yes

Note that *engineering* problems, based on how we've defined them, cannot be solved by a biological resolution because they were concerned with physico-chemical or thermodynamic aspects of the substrates/products. **There is one exception here. We might consider improvement of equilibrium reactions by coupling product removal to a second reaction to be a biological solution if we use a second enzyme. However, if we simply add a free enzyme, we haven't really changed our initial catalyst, so our definition of a biological solution isn't valid. It is valid if we alter a whole cell catalyst.* Likewise, some biological problems cannot be resolved with an engineering solution. All biological problems, however, can potentially be addressed with biological solutions.

The biological challenges we've discussed so far include the following:

- Enzyme instability (could be as a function of reaction conditions, like pH)
- Substrate inhibition/toxicity
- Product inhibition/toxicity
- By-product formation

The following enzyme-specific issues were also on your handout as having only biological solutions:

- Low enzyme activity
- Wrong or limited substrate specificity
- Poor selectivity

Each of these problems can be addressed by altering the catalyst, whether it is a free enzyme or a whole cell.

1. Characteristics of biological solutions

Recall that we defined a biological problem as one being specifically related to the nature of the catalyst. For the purposes of this discussion, we'll assume that simply screening for a better catalyst is not an option to our biological problems. So, once we've exhausted engineering solutions, we can turn to biological solutions for a remedy. Our objective now is to alter the existing catalyst to make it better. We can do this for either purified enzymes or whole cells, and we can use a random or a rational approach (see chart below). In this case, a random approach is defined as one in which you make many different versions of your base catalyst and then use screening in order to find an improved variant. A rational approach implies that a targeted, specific change was made to the catalyst in order to obtain an improved enzyme or whole cell. We can define these random vs rational approaches for biocatalysis in the following way (refer to Biological Solutions Handout):

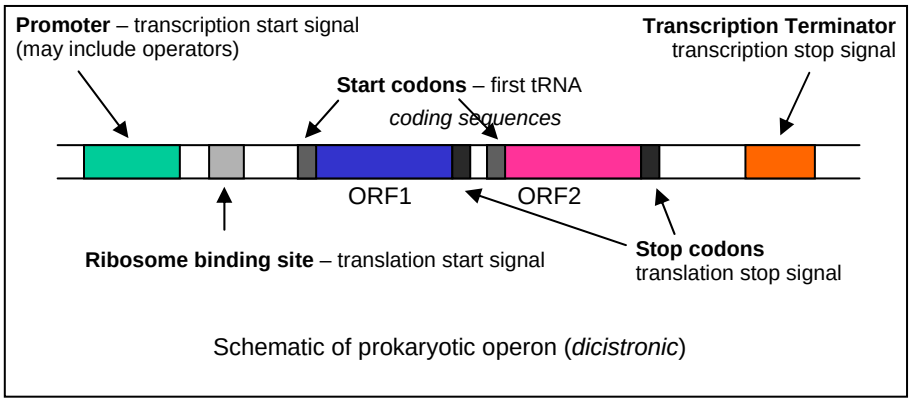
	"Random" Approach	**"Rational" Approach**
Whole Cell Catalyst	Classical Mutagenesis	Cloning or Metabolic Engineering
Purified Enzyme	Directed Evolution	Site-Directed Mutagenesis

The arrow in your handout is meant to convey the historical hierarchy of these techniques. You can also see that it follows along with the amount of genetic information that is required to implement the solutions. It's no coincidence that method chronology and required information content track so well! In general, those solutions that require more information will require more time to implement as well. Therefore, your ability to implement biological solutions to your design challenges will depend in part on how much time to have to complete your process design.

2. Review of Molecular Biology

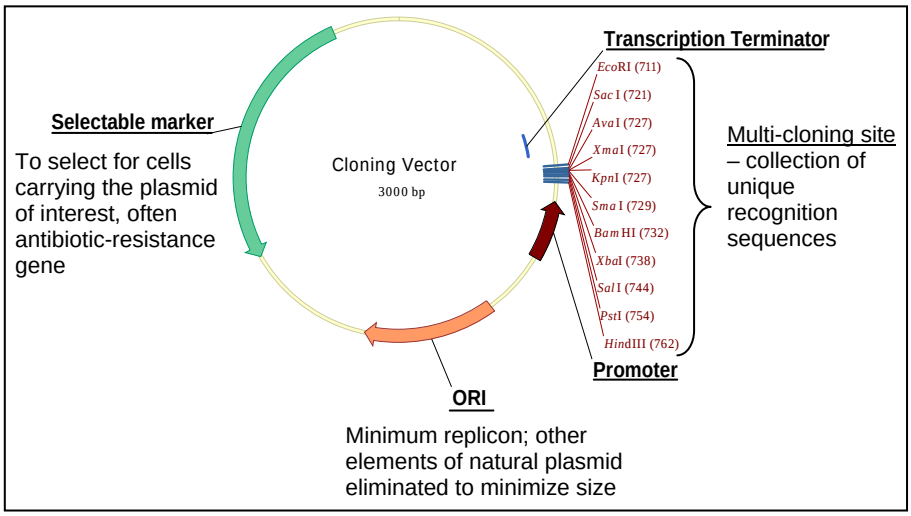
Recall the central dogma of molecular biology, that DNA goes to RNA and on to protein. Thus, by manipulating the DNA, we can influence the resulting protein. Recall as well that *how* and *where* the replication, transcription, and translation processes work differ between prokaryotes and eukaryotes. So the strategies that we'd use for manipulating these systems would differ as well. There are two extremes to the use of this dogma for genetic engineering: (1) adding new genes or activities, (2) deleting existing genes or activities. To address process design challenges, we may need to do either or both of these as well as manipulate new or existing genes.

In the case of biological manipulation of a whole cell system, we want to focus on gene expression. The target activity (*desired or undesired*) may be positively- (inducible), negatively- (catabolite repression) or un-controlled. Most control will be at the level of transcription. So, if you want to alter the expression of a gene (on or off), while not changing the sequence of the protein, you can create a change in the promoter region. If you're looking at a purified enzyme process, you'll want to use biological solutions to create an enzyme that has better properties than the original one. In this case, you'll want to modify the actual coding sequence (the open reading frame).



Schematic of prokaryotic operon (*dicistronic*)

To add new genes to a whole cell system, or to make a new protein, you need some way to introduce the DNA encoding for the new proteins into the original/host cell (whole cell system) or a production cell (purified enzyme system). The central dogma applies to DNA sequences that are part of the genome and elsewhere. We typically use the "elsewhere" option, and the most common tool for this is a plasmid. Plasmids are extrachromosomal, covalently continuous (circular) double-stranded DNA molecules. They are found mostly in bacteria, Gram-negative and Gram-positive, and are also available for yeast. (Remember that yeast and bacteria are our two most common choices for whole cell catalysts.) Plasmids are much smaller than the genome and more easily manipulated. Key features of plasmids are indicated on the diagram below.



We'll talk more about specific molecular biology tools required to achieve biological solutions using each of the approaches listed on your handout.

3. Classical mutagenesis of whole cell catalysts (Random Approach)

Classical mutagenesis is the oldest approach for improving catalyst activity and it requires the least amount of information. The basic strategy is to expose whole cells to a mutagen, *eg*, a chemical or UV light, and then screen for improvement in the desired activity. You don't need any information about the specific enzyme you're targeting; you only need to be able to grow and mutagenize the cells, and assay for your desired function. As an example, consider once again the Buckland reference for conversion of indene to indandiol. The researchers sought a biological solution to the problem of toluene requirement for induction of the desired TDO enzyme activity. This activity was tied to the ability of *Pseudomonas putida* to grow using toluene as a carbon source, and it relied on the enzyme activity to produce a colored compound, indigo. So the screen consisted of looking for cells grown without toluene that could make indigo. In this way, a mutant that was inducible by indene was identified, as well as a mutant that did not make TDO. The latter was subsequently screened for the ability to grow on toluene again, and a constitutive producer of TDO was identified. Given that TDO expression was on, then off, then on again, this likely implies that the mutations were in the promoter/operator region of the expression cassette and that the protein sequence itself was not altered. Note that although information about the TDO sequence was available for this system, it was not exploited for the random approach. This same experiment could have been done without any information on the system at all.

You can also employ the principles of classical mutagenesis without adding a mutagen and in this way, screen for strains with improved performance. This method relies on adaptability of cultures (*ie*, evolution on a short time scale) to a given growth environment. By using a chemostat – a CSTR for cell growth – you can make step changes in the culture environment and see how the cells adapt to a new steady-state. Staying with the indene example, the *Rhodococcus* I24 strain was chosen for further development work. By growing in a chemostat and subjecting the cells to a higher indene concentration, a strain with higher indene utilization was obtained. The yield of the desired product increased from 25% to 55%, and the proposed reason for this increase was the loss of two by-product pathways. Note that no genetic information was used to obtain these results either. Both of these examples show that you can get a lot of improvement with only a little information.

4. Cloning of enzyme activities/metabolic engineering of whole cells (Rational Approach)

The rational approach to altering a whole cell catalyst involves a more targeted methodology and subsequently requires a bit more biological information. There are two ways to rationally modify whole cells:

- (1) by cloning the target activity for expression in a new cell
- (2) by engineering the existing cell for better performance

The biological challenges we are most likely to be addressing by these approaches are:

- toxicity
- by-product formation
- We can also address the challenge of **low enzyme activity** by increasing expression of the target enzyme in order to get more of the desired protein (activity) per cell.

Let's first consider the case of cloning. Our objective here is to find the stretch of DNA in the host chromosome that corresponds to the enzyme of interest and them remove or copy that sequence so that it can be transferred to a new host. Why would we want a new host? Consider again Whole Cell Case Study #1 to answer this question (Courtesy of Merck & Co., Inc. Used with permission.). Recall that two by-products were identified in addition to the desired allylic alcohol. Through further analysis, it was determined that the enone to saturated ketone reaction was irreversible. This confirmed that any substrate going to saturated ketone would be a yield loss. Selective inhibition was used to improve the yield, but there were other biological challenges in this process as well (refer to p. 2 of handout):

- Volumetric productivity low due to substrate toxicity (cell) at high concentration (>2-3 g/L)
- Also evidence of product inhibition/toxicity (cell)

Note that cellular toxicity was important in this case because the cell was supplying co-factor. Once the cells died, co-factor recycling also died and the reaction stopped. It then seemed reasonable that a recombinant biocatalyst could be better than the original cell:

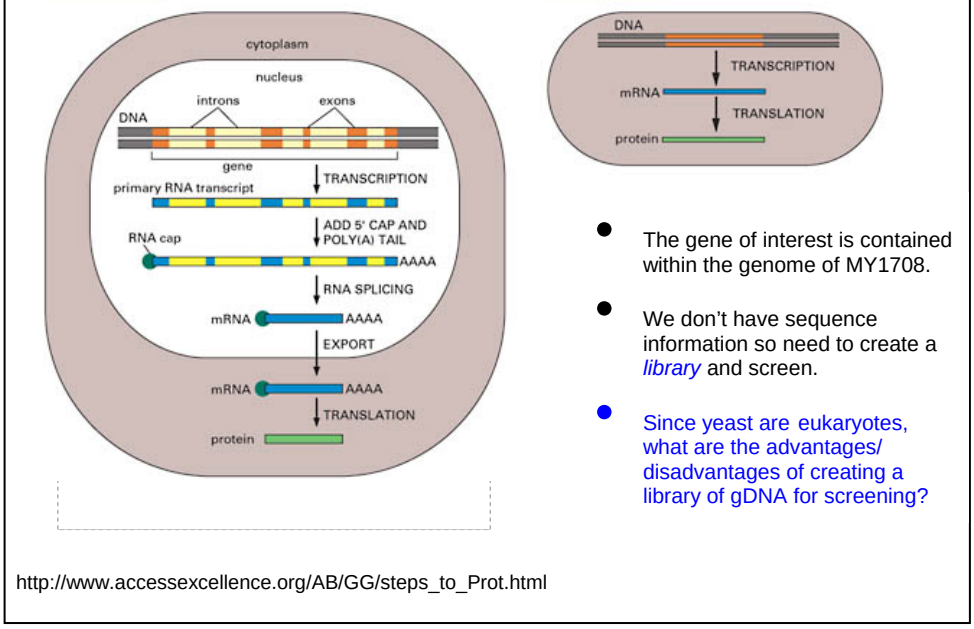
- Different host may not be subject to toxicity
- Ability to over-express enzyme means greater specific yields (relative to biomass) even if toxicity remains
- Could further reduce side reactions if additional reductase activity not present

So, the decision was made to attempt to clone the enone reductase activity for expression in a new cell. Because the starting source cell was a yeast, *Candida chilensis*, it was decided that the new host strain should also be a yeast. *Saccharomyces cerevisiae* (Baker's yeast) was chosen since it is well-characterized and molecular biology methods are readily available for cloning in this cell. A commercially available yeast plasmid was used as the DNA vector.

This system had one main drawback for cloning purposes – the DNA sequence of the enzyme was unknown. So while there was more biological/biochemical information (like the reversibility or irreversibility of the enzymes in the pathway), there was no genetic information available. In the absence of such information, you have to fall back on a semi-random approach of building a DNA library and then screening for the desired activity. For bacteria, this library would be formed as follows:

- Isolate genomic DNA
- Chop it up!
- Insert into plasmid molecules
- Put plasmids into (transform) cells
- Screen assuming 1 cell = 1 plasmid = 1 stretch of genomic DNA

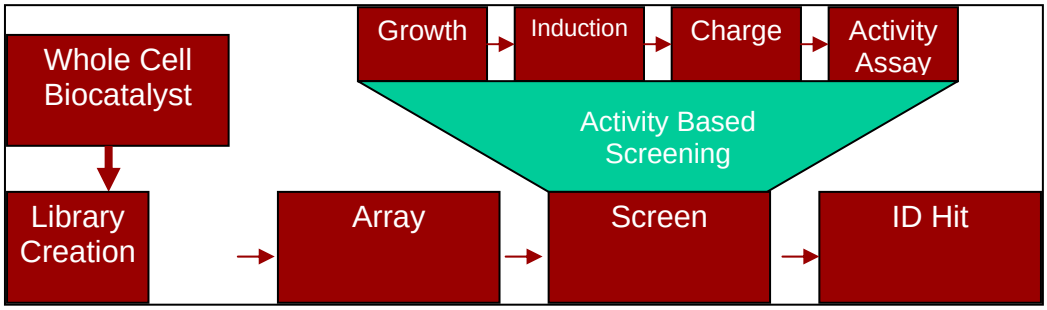
This strategy will almost certainly fail for a yeast strain. *Why is this?* Recall that yeast are eukaryotic, so the gDNA has both introns and exons (diagram below). If you create a gDNA library, you will need larger fragments of DNA to get the full gene, and more importantly, you are now relying on the new host to splice the gene correctly. Taking from yeast to yeast means it's more likely that the splicing would be correct, but you're taking a big change in this case. So, what's the alternative?



Instead, you can use a cDNA library, created from the spliced mRNA that will be present in the cytoplasm. The same principles apply as with a bacterial gDNA library except that you make cDNA from mRNA and use that as the DNA source for your plasmid library. In order to get mRNA, you will need cells grown under conditions that result in high activity. *Why?* Remember that the growth phase is the enzyme production phase. For a genomic DNA library, you would get representative fragments encompassing all genes, whether protein is being made from them or not. With a cDNA library, you're starting point is the mRNA, which is representative of only those genes that are transcribed.

Once you have created your library you need to find the "needle in the haystack," *ie*, the one to few clones in your library with the right properties. This is where the "random" aspects of this approach come in, namely, that we need to screen the library to uncover a clone that has activity against the target substrate. The best screens are ones like the

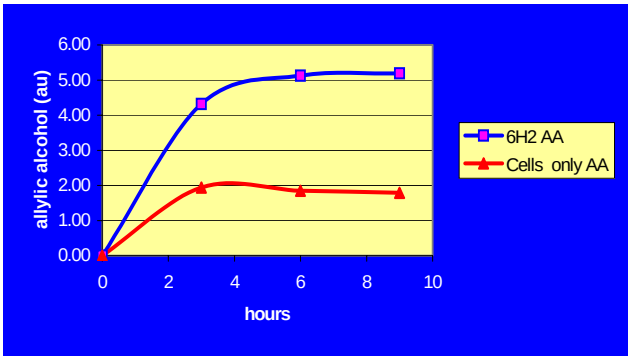
indene/indigo example, where the screen was for blue cells. This case was not so blessed, so the screen had to be for the actual product of interest. The screening methodology is shown in your handout (diagram below).



For this type of approach, you also need to know how many clones to screen. We'll make the following assumptions to answer this question.

- Assume genome of ~6,000 genes (*S. cerevisiae*)
- To achieve 95% confidence interval that each gene is screened once, need 4-fold coverage
- Total screen = 24,000 clones (250 96-well plates)

The "fast" HPLC method used here took 2 minutes per sample. At this rate, it would have taken about a month (33.3 days) to screen 24,000 clones if the instrument ran without stopping! What you are looking for in this screen is a sample that is better with plasmid DNA containing a cDNA fragment than without it. One lead candidate had this property, named 6H2.



But now that you have a candidate clone, you need more biological information. In a cloning exercise, you want to know the genetic sequence, so if you don't start with this knowledge, you want to end with it. If you know the sequence, you can search databases to see what the gene "looks like," *ie*, does it have a sequence that is consistent with the activity displayed, *eg*, does it look like other dehydrogenases (reductases)? If the answer is yes, you're off to a great start. If the answer is no, you may need to worry a bit. In this case, the answer was no, and in fact, the gene was consistent with a mitochondrial carrier protein. An MCP is a transporter protein and helps molecules get from the cytoplasm

into the mitochondria. (Remember that eukaryotes have organelles, or compartments, within the cell.) A transporter protein is not likely to be an enzyme (*ie*, perform a chemical conversion). So, more information is needed here. Additional studies were done and the following facts about the system were uncovered (p. 5 of the handout):

- The control host (no cDNA) had activity similar to clone 6H2 upon lysis (*ie*, when walls are removed).
- The original *C. chilensis* MY1708 system showed NADPH-dependence for conversion of enone to allylic alcohol, while 6H2 was NADH-dependent.
- The 6H2 lysate system also had low EE, ~30% compared to >95% for MY1708.

More news was not good news. The information here collectively suggested that the cloned DNA fragment does not encode for an enzyme, but rather encodes for a transporter that helps the substrate get into the cell or from the cytoplasm into another organelle where it is converted to allylic alcohol but with poor selectivity. This is sadly where the story ends, and it demonstrates that even with a clear target, biological solutions are more challenging than engineering ones, and may be much more difficult to implement.

Let's look again at Whole Cell Case Study #2, for the selective conversion of indene to indandiol (p. 6 of the handout). Remember that the first strain used was *Pseudomonas putida*, and that the target enzyme was toluene dioxygenase. As an alternative to the random approach of classical mutagenesis, a rational cloning approach was used to eliminate the need for toluene induction of the TDO activity. Let's answer the following questions from the handout:

- **How much genetic information was known?** The sequence of this gene was already known, so rather than creating a library, the desired gene could be PCR-amplified directly from genomic DNA of the host. *Do we need to worry about cDNA here? No, because Pseudomonas is a prokaryote.*
- **Why *E. coli*?** – Because the host is "naïve" meaning it does not already metabolize indene so no need to worry about side reactions.
- **What was the result, and was it good or bad?** This system behaved pretty much like the *P. putida* system. When only the ABC gene components were present, the EE was about 30%, the same as was observed in the early stages with the *P. putida* system. This meant that the intrinsic EE of that enzyme was ~30%. By including TdoD, the EE increased substantially but it was accompanied by a drop in total cis-indandiol yield since the undesired enantiomer was removed.

This example demonstrates that simply moving an enzyme from one host to another doesn't change it's inherent properties, so while this system was indeed free from toluene-induction, it was still a poorly performing process. The take-home message from both of these examples is that this approach can be powerful if it works, BUT it may take some time to work out all the "bugs" in the system. This also emphasizes that an engineering solution, if effective, is likely to be achieved more rapidly.

Let's now talk about metabolic engineering – Remember that we've defined metabolic engineering as the targeted alteration of the starting whole cell catalyst for improved performance. This differs from cloning, where we want to *remove* the target activity for expression in a different host. Continuing with Whole Cell Case Study #2, a *Rhodococcus* strain was sought that would have higher initial EE (*ie,* did not require a dynamic upgrade) and was found. The last strain mentioned in the Buckland paper, I24, was studied here at MIT by Greg Stephanopoulos in this department and Tony Sinskey in the Biology department. Where we left off previously, strain KY1 had double the yield of desired product (55% vs. 25%). In this case, two pathways (to indenol and one cis-indandiol) were lost just by selection with higher indene exposure. You can think of this as "accidental" metabolic engineering. For "real" metabolic engineering, however, you need to target specific pathways.

Part of metabolic engineering is to actually figure out what those targets should be. Starting with strain KY1 (which had also lost the ability to further metabolize the trans-indandiol), studies were conducted to figure out where the substrate was going. It was discovered that >95% was going to the indene oxide compound and that this was non-enzymatically hydrolyzed to the two indandiol enantiomers (both are desired products). This hydrolysis was slow and 2/3 of the indandiol was being further metabolized. Based on this analysis, it was determined that if the oxide hydrolysis rate could be enhanced, you could obtain higher yields of the indandiols. This time, a cloning exercise was done to take a known epoxide hydrolase from a different *Rhodococcus* strain and introduce it into I24. This new strain had an increase in product yield from ~55% to ~75%. Finally, a pH optimization (an engineering solution!) was done that further increased the yield to ~95%. This is a successful example of metabolic engineering, and it demonstrates that while you do need biological information for a rational solution, you can gather some of this information as you go.

5. Site-directed mutagenesis of purified enzymes (Rational Approach)

The cloning approach to whole cell catalyst optimization leads to a known DNA sequence for the enzyme of interest. Once the DNA sequence is known, it's fairly easy to mass-produce the enzyme (over-express it) in a host cell and obtain a purified enzyme that can then be used for biocatalysis if desired. Knowing the sequence also opens up the possibility of further optimization by changing the amino acid sequence in order to change the catalyst behavior. As we continue along our solutions options chronologically, the rational approach of site-directed mutagenesis (SDM) is the next methodology to examine. SDM is most often used to change **the specificity** of an enzyme to accept more substrates or the **tolerance** of an enzyme to environmental conditions (pH, temp, solvent concentration). This approach involves changing specific amino acids to one or more other amino acids and testing the resulting proteins for changes in properties.

The key to SDM is to know *which amino acids* to modify. Thus, in addition to knowing the DNA sequence, and by extension the amino acid sequence, you need more information about how structure is related to function. For example:

- Which amino acids contact the substrate in the active site?
- Which amino acids are exposed on the surface?

Usually, you gain this information from a crystal structure of the protein. Unfortunately, crystal structures are hard to come by and there are far fewer of them available than the number of enzymes that have been sequenced. So while SDM actually arose as a technique before directed evolution, it actually requires the greatest amount of biological information to successfully implement.

As an example, let's consider the case in your handout. This is not a specific case of biocatalysis, but the class of enzymes studied here are ones that have been used for industrial biocatalysis (can refer back to the Survey of Biocatalytic Reactions from the second day of class for examples). In this paper, the authors wanted to change the specificity of phenylalanine dehydrogenase (PheDH). It has a preference for aromatic amino acids, Phe and Tyr. The enzyme LeuDH has a preference for aliphatic residues. The authors wanted to change the specificity of PheDH towards residues that are preferred by LeuDH. PheDH's and LeuDH's have high sequence homology (~50% identity), so it makes sense that the amino acid sequences would "map" structure to function. If the amino acid residues that were known to form the active site in PheDH could be changed to the corresponding aa's from LeuDH, then the specificity would likely change as well. The question here was, *which amino acids should be targeted*?

No crystal structures were available for PheDH or LeuDH, but a structure was available for GluDH. This enzyme belongs to the same superfamily of amino acid dehydrogenases, but the identity is only ~20% to PheDH or LeuDH. Two amino acid residues, Ala163 and Val377, were known to contact the substrate in GluDH. The authors used a technique called homology-based modeling to identify the corresponding amino acids in PheDH and LeuDH. They found that the same two amino acids were present at the corresponding position in LeuDH, but that they were replaced by Gly and Leu respectively. So, the strategy was to change these to Ala and Val in order to change the substrate specificity (based on the universal code) so that it was more like LeuDH. The results indicate that the strategy worked and activity against aliphatic side chains was substantially increased. Activity against Phe remained but was significantly reduced, especially in the double mutant. The results indicate that SDM can be applied very successfully, but the major drawback is the (vast!) amount of information required.

6. Directed evolution of purified enzymes (Random Approach)

The last technique we'll discuss is the "random approach" of directed evolution. This technique requires a cloned gene as well; however, it does not require the same amount of biochemical information as SDM. In fact, directed evolution exploits the fact that we probably know much less about "well-characterized" enzymes than we think we do! Mutations have naturally arisen in protein sequences and have resulted in enzymes with better or different activities. But the natural mutation rate is very small (*eg*, DNA polymerase from *E. coli* has an error rate of $1:10^6$ bp). Directed evolution is a methodology that purposefully introduces mutations into the DNA sequence (and the

related protein sequence) at a much faster rate than natural evolution. These mutations will be introduced throughout the protein and not just in the active site. One common way to do this is to use "error-prone" PCR. This involves the use of a DNA polymerase with a much higher error rate than $1:10^6$. Remember that we've defined random approaches as ones that require screening or selection to identify beneficial mutations, and that's the key here as well. Your PCR reaction will give a variety of different sequences, so you'll need to screen to identify altered proteins with better characteristics.

- *One other note about DE* – Because this technique alters all regions of the protein, it can be used to target almost any property of the enzyme, *ie*, all of our biological problems. This doesn't mean it will work, but it can be tried.

Let's look at our final example, again courtesy of Merck & Co., Inc. (Used with permission.) (p. 4 of handout). The chemical shown here is an intermediate in the production of a compound for the treatment of benign prostate enlargement, and the reaction is a resolution. The *optimized* 2nd delivery process was still slow (12 days to completion) and required a large amount of enzyme (50% by weight). So the final solution employed was directed evolution to try to improve the enzyme activity. Page 6 of your handout shows an example Screening Result. Since higher activity was the target criterion here, the researchers looked for altered sequences that would convert more substrate in a given amount of time. The screening results graph shows the (relative) range of product amounts obtained after a round of directed evolution. Note that directed evolution can be done in an iterative fashion. The best performer from one round can be used as the template for round two. After two rounds here, activity was increased five-fold (p.7 of handout).

The final step in directed evolution is to characterize the products. The last page (p. 8) of your handout shows the mutations obtained for 3 mutants in Round 1 and the best mutant from Round 2. In the first round, the number of mutations ranged from 1 to 3, and the variant with 2 mutations was carried forward to Round 2. The Round 2 Positive contained an additional mutation. Note that a total of 5 positions were found to be altered here, but only two were known from the crystal structure to contact substrate in the active site. In the final variant, two out of the three mutations were in the active site. Because the substrate here is a non-natural one, you might think that increasing the activity is simply a matter of changing the active site to better accommodate the substrate, so positions 62 and 224 might have been altered in a rational approach. However, changing aa43 was also beneficial, and this would likely not have been an obvious target from the structure. Therefore, directed evolution has the advantage of being able to give you changes that you might have targeted through SDM but also giving you other ones that you would not have targeted.

7. Final thoughts

I've tried to give you an overview of biocatalysis, and I hope that you can see what a powerful tool it can be. As an engineer, you are concerned with optimizing processes,

and you have a variety of tools available to do so. These can be chemical or biological in nature. The key is to use everything you have–but only when you need it–and use it well!