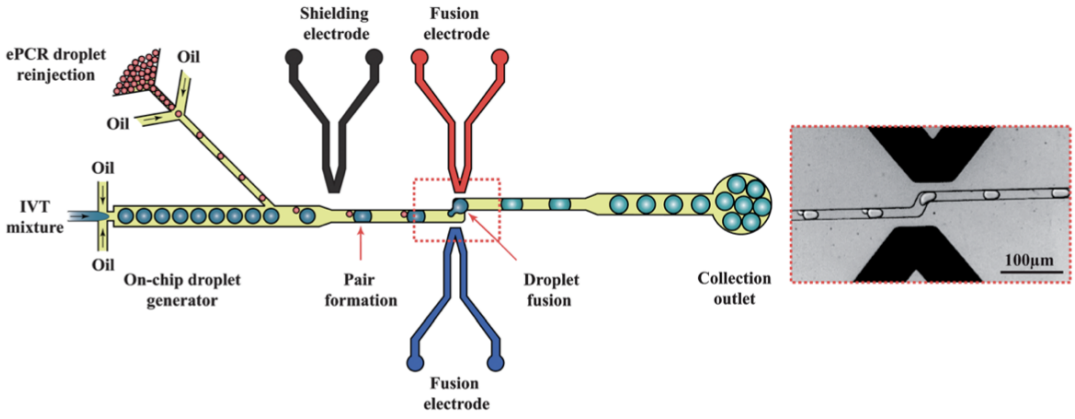
2018 ChemE 355
Problem Set 4
Due Tuesday, June 5th in class

Problem 1: Protein engineering using droplet microfluidics

In class we discussed the benefits and challenges of evolving a new protein in vivo. We also discussed ways to encapsulate proteins and the DNA that encodes them to improve the throughput of in vitro protein engineering. A paper by Griffiths et al (Lab on a Chip, 2012) describes one such approach to in vitro protein engineering using cell free protein synthesis within droplets and illustrates some of the complexities involved in making this work. The overarching goal is to develop an ultra-highthroughput method using both in vitro expression and assay of a large library of enzyme variants. As a proof of concept experiment, they worked with the enzyme beta-galactosidase LacZ, which can be directly assayed using fluorescence (LacZ can hydrolyze the substrate fluorescein-di-β-D-galactopyranoside to generate the fluorescent product fluorescein). To test if in vitro protein expression and assay would work, they expressed the enzyme lacZ and an inactive variant, ΔlacZ that contained a frame-shift mutation within the middle of the sequence. They then demonstrated that droplets (20 pL each) containing the expressed protein could be sorted by fluorescence to isolate only the active beta-galactosidase and not the inactive variant as an important step towards developing a high throughput in vitro complement to in vivo protein engineering.

Overview of workflow to generate droplets that will then be sorted based on fluorescence: (IVT mixture provides components for transcription and translation and protein assay; this mixture will be fused to droplets containing PCR products as shown)



For the emulsion PCR step, the authors determine that a concentration λ of 0.15 molecules of template DNA per droplet on average is ideal for their system. This results in ~12% of droplets containing one molecule of template, and <1% of droplets containing >1 template molecule.

(a) Why is it important to limit the number of droplets that have >1 template molecule? What would happen if a significant % of droplets contained 2 template molecules? **(2 pts)**

At a high level, we simply can't correlate protein activity directly to DNA sequence if we have multiple DNA templates in a single droplet!

Further elaboration (not required to answer the question):
The goal of this proof of concept is to show that you can totally separate the two types of DNA based on functionality. Getting two strands into a single droplet to start means that you can not separate based on sequence. This can result in having inactive droplets in the positive pool or active droplets in the negative pool (depending on how thresholds are set) and a higher false positive/negative rate.

(b) What would happen to the signal in the beta-galactosidase reaction if one template molecule each for lacZ and ΔlacZ are encapsulated in a droplet? How would this signal be interpreted in terms of protein function? **(2 pts)**

The droplets have a weaker fluorescence signal due to having a combination of lacZ and ΔlacZ enzymes within them.

Further elaboration (not required to answer the question):
After amplification, half of the DNA pool will be lacZ and the other half will be ΔlacZ. After transcription, this ratio will be maintained. And after translation, it will again be maintained. Therefore the protein pool will contain half functional and half non-functional protein. The signal will then be between the completely functional or completely nonfunctional droplets.

(c) After 32 rounds of PCR, how many total strands of DNA should they theoretically have observed per droplet if they started from one double stranded molecule of template? (Doesn't need to be exact, order of magnitude answer is ok) **(2 pts)**

2^n+1 = 2^33

Accept anything from $8.6*10^8$ and $8.6*10^{10}$. Order of magnitude. 2^33 or 2^32 are fine.

(d) In practice, the authors observed 30,000 copies of each gene for droplets containing 1 molecule of DNA template. What is one reason that could account for the discrepancy between this number and your answer to part b? **(2 pts)**

The discrepancy is caused by the droplets running out of material for the PCR reaction

(e) Does the strategy describe a screen or selection? Please explain in one or two sentences what the difference is between a screen and a selection, and which one better describes this technique. **(2 pts)**

This is a screen. Screens primarily rely on different variations possessing a characteristic that allows for identification and usually sorting of the variants. Selections describe the case where you are only gathering information about sequences or clones that are successful with regard to your assay. Since this work sorts drops based on fluorescence, and all droplets are analyzed, it is a screen, even though the following step of discarding non-fluorescent cells resembles a selection.

(f) Why go through the trouble of doing transcription and translation in the droplet? Why not just do transcription/translation in bulk solution and then encapsulate protein molecules? Assume for simplicity that you can get enough signal from one molecule of protein to enable droplet sorting. **(2 pts)**

Because you need to know the sequence of the enzyme after the screen is complete – so DNA and enzyme must be kept together.

(g) Imagine this approach were used in a paper entitled, "An engineered variant of LacZ that has >1000x the activity of the wild-type enzyme". In this hypothetical paper, error-prone PCR is used to generate a library of LacZ mutants, which were transcribed and translated to protein in droplets and then sorted. What is the next step after the final round of screening? **(2 pts)**

Choose the drops that fluoresce the most, collect DNA and amplify it (either by putting it into E coli, or with in vitro PCR amplification). This DNA will then be sequenced to determine the sequence of successful enzymes.

(h) In order to produce protein, each droplet containing DNA after PCR is then fused with a second droplet containing transcription and translation reagents (see figure, above). Given 30,000 copies of DNA/droplet, a transcription rate of 48 nucleotides/sec, an extra long transcript half life of 10 min, and a translation rate of 16 AA/sec, what is the frequency of translation initiation in the cell-free system for lacZ (1023 residues)? **(5 pts)**

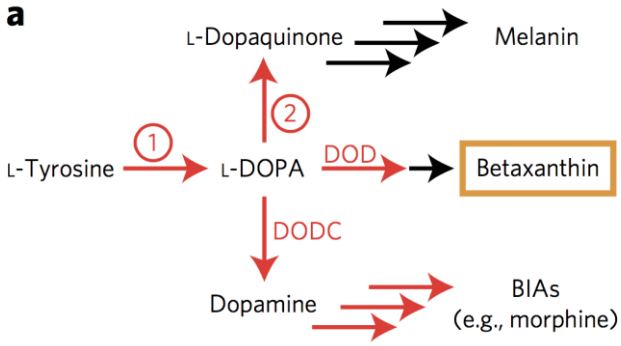
Exp(-k*600) = 0.5 → k = ln(2)/600
dmRNA/dt = 48 nuc/sec/DNA * 30000 DNA*1 mRNA/(1023*3 nuc) – k *[mRNA] = 0
[mRNA] = 4.06e5
initiation rate for one mRNA = 100 nuc bt rib/48 nuc/sec = 2.1 sec between initiations ➔ .48 init/sec/mRNA
total initiation frequency = 0.48 init/sec/mRNA * 4.06e5 mRNA = 194880 init/sec total

(i) For the drops to register a fluorescent signal, they need a concentration of at least 100 uM of fluorescein. Assuming constant linear production of lacZ, how long do the scientists have to incubate the drops to create enough flourescein to trigger the detector given that lacZ has a kcat of 15 flourescein/sec/lacZ (Avogadros # = 6.022e23) **(5 pts)**

20e-12 L * 100e-6 mol fl/L = 2e-15 mol fl * 6.022e23 = 1.2e9 fluorescein
prod rate of lacZ = 16AA/sec/ribosome*1023*3/100 rib/mRNA * 4.06e5 mRNA * 1 lacZ/1023 AA = 194880 lacZ/sec
amount of lacZ is a function of time = 194880*t(sec)
1.2e9 fluorescein = integral[15 fl/sec/lacZ * 194880*t (lacZ), t, 0, t*]
solving gives t* = 29.0 sec

Problem 2: Protein engineering for morphine biosynthesis

In a 2015 paper in Nature Chemical Biology about engineering the early steps of the morphine pathway into yeast, Dueber and coworkers built a library of 200,000 CYP76AD1 mutants in order to enhance enzyme-catalyzed conversion of tyrosine to L-DOPA (the first step in the morphine pathway) and also limit a second oxidation also catalyzed by the WT enzyme that converts L-DOPA to the L-Dopaquinone (en route to melanin – see figure below). The paper is posted on canvas, but is not required to answer the questions.



(a) In order to search for the desired activity of a CYP76AD1 variant, described above, the authors developed an in vivo biosensor that provided a direct, fluorescent readout for L-DOPA production. In addition, the authors chose to develop a second assay that provides an indication of the combined L-DOPA and L-Dopaquinone pools by production of a violet pigment. What additional information is given in this second, less specific assay? **(2 pts)**

The first assay for L-DOPA production can be used as a positive screen to identify clones that have enhanced production of L-DOPA. The second assay can be used, in combination with the first to carry out a negative screen for clones that lack the undesired DOPA oxidase activity. Even if a variant produces high levels of L-DOPA, if much of that product is funneled to melanin (through L-Dopaquinone) it will not be useful for production of morphine.

(b) The authors note that 6 of the 17 clones that they characterized with improved L-DOPA production contained the same F309L missense mutation, and that among these, there were two distinct codon changes. Imagine that one has the Phe codon UUU mutated to UUA (Leu), and the other has UUU mutated to CUU (a different Leu codon). What is the significance of this finding? List two reasons why amino acid changes in a given enzyme sequence could lead to accumulation of more product. **(4 pts)**

The result means that the phenylalanine to leucine mutation occurred by two different independent mutation events and is likely very beneficial for L-DOPA production. These mutations could result in an improved kcat for the Tyrosine to L-DOPA reaction, or could result in more expression of protein.

(c) DNA shuffling and another round of error-prone PCR is then performed using the six mutants with the largest improvements in fluorescence (indicating the highest yield of DOPA). What is the theory behind DNA shuffling, and why is it superior to the simpler alternative of just another round of error-prone PCR of the mutant sequence with the highest activity? **(2 pts)**

DNA shuffling allows for beneficial mutations that have evolved independently to be combined (to test for synergistic effects), and deleterious or neutral mutations to be lost.

(d) What is a technique the authors could have used to identify deleterious mutations in a single clone that might have accumulated during the process of error-prone PCR? Describe the major steps involved, and how deleterious mutations would be identified. **(4pts)**
The authors could have performed backcrossing of the mutated sequence with the parent WT sequence. This involves shearing the DNA, PCR amplification, expression in the host and screening for activity, and finally sequencing a panel the top-performing clones. Only beneficial mutations should persist in all of the clones sequenced. Deleterious mutations should be lost, and neutral mutations should only be present in a fraction of the sequences, if at all.

(e) The following table taken from the supplemental information lists the mutations found in the six variants that were used to undergo DNA shuffling in the paper.

	Mutations				
Mutant #1	*87A>C*	L141I			
Mutant #2	*123T>A*	F309L	E465D		
Mutant #3	D2E	*150G>A*	*327C>T*	Y380H	
Mutant #4	*9T>C*	W13L	*1236T>C*	*1281T>C*	
Mutant #5	*180T>A*	F127L	*576G>A*	S232T	*714G>A*
Mutant #6	*684C>T*	F309L			

After performing a round of DNA shuffling with these six mutants, the authors determined (by sequencing active clones) that the combination of mutations F309L and W13L are most beneficial for enzyme activity. If we were to recombine the mutant sequences #1-6 using DNA shuffling (but without the error-prone PCR step), what is the smallest library size needed in order to obtain 20 clones containing these two mutations? Assume equivalent starting amounts of each mutant sequence are used. **(4pts)**

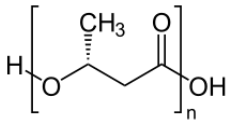
The probability of incorporating F309L is 1/3, and the probability of incorporating W13L is 1/6. Consequently, the probability of incorporating both into a given clone would be 1/18 = (1/3)*(1/6).

20/x = 1/18 → x = 360 clones in the library

(f) Aro4P catalyzes the first step in the shikimate pathway in yeast (recall the shikimate pathway is the source of all aromatic amino acids) and is known to be quite sensitive to tyrosine feedback inhibition. Luckily a variant Aro4P*FBR* had previously been identified that does not suffer from this problem. What is feedback inhibition, and why is it problematic for Aro4P in the context of L-DOPA production in yeast? **(2 pts)**

Feedback inhibition is a description for allosteric regulation of an enzyme by it's product, and has evolved such that primary metabolite levels can be tightly controlled. The precursor to L-DOPA is tyrosine, an aromatic amino acid that comes from the shikimate pathway. If high levels of tyrosine limit flux through the shikimate pathway then substrate might be limiting for CYP76AD1. (2 pts)

Problem 3: Production of the bacterial plastic polyhydroxybutryate



Poly-3-hydroxybutyrate (PHB), structure above, is a biodegradable thermoplastic accumulated intracellularly by many microorganisms under unfavorable growth conditions. *Azotobacter chroococcum* is being investigated for commercial PHB production using cheap soluble starch as the raw material and ammonia as the nitrogen source. Synthesis of PHB is observed to be growth associated with maximum production occurring when the culture is provided with limited oxygen. During steady-state continuous culture of *A. chroococcum*, the concentration of PHB in the cells is 44% w/w and the respiratory coefficient is 1.3. From elemental analysis, *A. chroococcum* biomass without PHB can be represented as $CH_2O_{0.5}N_{0.25}$. The monomeric unit for starch is $C_6H_{10}O_5$; $C_4H_6O_2$ is the monomeric unit for PHB.

(a) Develop an empirical reaction equation for PHB production and cell growth. PHB can be considered a separate product of the culture even though it is not excreted from the biomass.

$C_6H_{10}O_5 + A\ O_2 + B\ NH_3 \rightarrow C\ CH_2O_{0.5}N_{0.25} + D\ CO_2 + E\ H_2O + F\ C_4H_6O_2$

We have 6 unknowns, so we need to generate 6 equations. We can begin with elemental balances to give us 4 equations.

N: B = 0.25*C

C: 6 = C + D + 4*F

O: 5 + 2*A = 0.5*C + 2*D + E + 2*F

H: 10 + 3*B = 2*C + 2*E + 6*F

We can generate one more equation using the respiration coefficient: 1.3 = A/B or

1.3*A = D

Lastly, we are told that the dry cell weight is 44% PHB (CO_2 bubbles off and H_2O is not considered dry cell weight). Note that the molecular mass of PHB monomer is 86.09 AU and the molecular mass of the cell unit is 25.53 AU. (-2 points if the 44% is interpreted as the ratio between PHB monomer and cell mass).

0.44 = 86.09*F/(86.09*F + 25.53*C) or 38*F + 11*C = 86.09*F or

11*C = 48*F

You can also consider the cell mass given to be the total cell mass including other elements. In Jim's lecture, he mentioned that only 92% of the cell mass is carbon, hydrogen, oxygen, or

nitrogen. Therefore, the cell mass given could be considered the TOTAL mass and the equation must be adjusted to account for the other elements present.

0.44 = 86.09*F/(86.09*F + 25.53*C/0.92) or 38*F + 12*C = 86.09*F or

12*C = 48*F

This gives us a system of equations with 6 equations and 6 unknowns, so we know it can be solved. The simplest method of solving this system is likely using matrices:

$$\begin{bmatrix} 0 & -1 & 0.25 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 4 \\ -2 & 0 & 0.5 & 2 & 1 & 2 \\ 0 & -3 & 2 & 0 & 2 & 6 \\ -1.3 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -11 & 0 & 0 & 48 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 6 \\ 5 \\ 10 \\ 0 \\ 0 \end{bmatrix} = [1.3 \quad 0.56 \quad 2.2 \quad 1.7 \quad 2.1 \quad 0.51]$$

Or with the 92% of cell mass adjustment:

$$\begin{bmatrix} 0 & -1 & 0.25 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 4 \\ -2 & 0 & 0.5 & 2 & 1 & 2 \\ 0 & -3 & 2 & 0 & 2 & 6 \\ -1.3 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & -12 & 0 & 0 & 48 \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ 6 \\ 5 \\ 10 \\ 0 \\ 0 \end{bmatrix} = [1.3 \quad 0.53 \quad 2.1 \quad 1.7 \quad 2.1 \quad 0.53]$$

Therefore, we can rewrite our original equation with all of the coefficients filled in!

$C_6H_{10}O_5 + 0.56\ NH_3 + 1.3\ O_2 \rightarrow 2.2\ CH_2O_{0.5}N_{0.25} + 1.7\ CO_2 + 2.1\ H_2O + 0.51\ C_4H_6O_2$

Or with the 92% of cell mass adjustment:

$C_6H_{10}O_5 + 0.53\ NH_3 + 1.3\ O_2 \rightarrow 2.1\ CH_2O_{0.5}N_{0.25} + 1.7\ CO_2 + 2.1\ H_2O + 0.53\ C_4H_6O_2$

(b) What is the yield of PHB-containing cells from starch in units of g g^{-1}?

Looking for the mass of only cells, not any other products.

$Y_{X/S} = M_{cells}/M_{substrate}$

$$Y_{x/s} = \frac{2.2 * 25.53}{1 * 162.14} = \mathbf{34.6}\% \ (33.1\% \ if \ 92\% \ adjustment \ is \ made)$$

Problem 4: Oxygen demand for production of recombinant protein in *E coli* (10 points)

Recombinant protein is produced by a genetically engineered strain of *Escherichia coli* during cell growth. The recombinant protein can be considered a product of cell culture even though it is not secreted from the cells; it is synthesized in addition to normal *E. coli* biomass. Ammonia is used as the nitrogen source for aerobic respiration of glucose. The recombinant protein has an overall formula of $CH_{1.55}O_{0.31}N_{0.25}$. The yield of biomass (excluding recombinant protein) from glucose is measured as 0.48 g g^{-1}; the yield of recombinant protein from glucose is about 20% of that for cells.

(a) How much ammonia is required? (4 points)

$$C_6H_{12}O_6 + a\ O_2 + b\ NH_3 \rightarrow c\ CH_{1.77}O_{0.49}N_{0.24} + d\ CO_2 + e\ H_2O + f\ CH_{1.55}O_{0.31}N_{0.25}$$

(1 point)

MW glucose = 180.2 g/mol; MW biomass = 25.0/0.92 (because C,H,N,O make 92% of E. coli) = 27.1 g/mol; MW recombinant protein = 22.03 g/mol

Yield of biomass = 0.48 g/g = Y_{XS}

$c = Y_{XS} * MW_{substrate} / MW_{cells}$ = 0.48 *g/g* * 180.2 *g/mol* / 27.1 *g/mol* = 3.19 (1 point)

Yield of recombinant protein = 0.2, but need to include yield of biomass, so

Y_{PS} = 0.20*0.48 = 0.096 g/g, and

$f = Y_{PS} * MW_{substrate} / MW_{product}$ = 0.096 *g/g* * 180.2 *g/mol* / 22.03 *g/mol* = 0.79 (1 point)

From the nitrogen balance: $b = 0.24*c + 0.25*f$ = 0.24*3.19 + 0.25*0.79 =

0.96 mol ammonia per mol glucose (1 point)

(if 0.2*0.48 ignored and used 0.2 instead, then f = 0.2*180.2/22.03 = 1.64, and b = 0.24*3.46+0.25*1.64 = **1.24**, 1 point taken off and no additional points taken off)
(if 25 g/mol is used for E. coli mass, then c = 3.46, f = 0.79, b = 0.24*3.46 + 0.25*0.79 = **1.03**, 1 point taken off no additional points afterwards)

(b) What is the oxygen demand? (3 points)

Determine oxygen demand using electron balance:

From class, degree of reduction of glucose relative to ammonia: γ_S = 4.00, and that of *E. coli* relative to ammonia: γ_B = 4.07.

The degree of reduction of the recombinant protein relative to ammonia:

$$\gamma_P = \frac{1*4+1.55*1-0.31*2-0.25*3}{1} = 4.18 \textit{ (1 point)}$$

From stoichiometric balance above, $w = 6$ for glucose and $j = 1$ for recombinant protein, then:

$a = ¼ * (w*\gamma_S - c*\gamma_B - f*j*\gamma_P) = ¼ * (6*4.00 - 3.19*4.07 - 0.79*1*4.18) =$

1.93 mol O_2 per mol glucose (1 point)

(if 0.2*0.48 ignored and used 0.2 instead, then f = 0.2*180.2/22.03 = 1.64, and b = 0.24*3.46+0.25*1.64 = 1.24, so a = 0.25*(6*4-3.46*4.07-1.64*1*4.18) = 0.77)

(if 25 g/mol is used for E. coli mass, then c = 3.46, f = 0.79, b = 0.24*3.19 + 0.25*0.79 = 0.963, a = 0.25*(6*4-3.46*4.07-0.79*1*4.18) = **1.65**)

Also can use system of equations without electron balance:

6 = c + d + f
12 = -3b + 1.77c +2e + 1.55f
6 = -2a +.49c +2d + e + 0.31f
0 = -b + .24c + .25f
0.48 = 0.139c
0.096 = 0.122f

	a	b	c	d	e	f	=
C	0	0	1	1	0	1	6
H	0	-3	1.77	0	2	1.55	12
O	-2	0	0.49	2	1	0.31	6
N	0	-1	0.24	0	0	0.25	0
Y_{XS}	0	0	0.1504	0	0	0	0.48
Y_{PS}	0	0	0	0	0	0.122	0.096

a = 1.9304
b = 0.9627
c = 3.1915
d = 2.0216
e = 4.0097
f = 0.7869

(c) If the biomass yield remains at 0.48 g g^{-1}, how much different are the ammonia and oxygen requirements for a wild-type strain of *E. coli* that is unable to synthesize recombinant protein?

No recombinant protein → f = 0 (1 point)

From N balance:

b = 0.24*c = 0.24*3.19 = 0.77 (E coli = 27.1)

b = 0.24*c = 0.24*3.46 = **0.83** (E coli = 25, no extra points off)

Thus 0.77 mol ammonia is required per mol glucose. (1 point)

From electron balance:

a = ¼ * (w*γ_S – c*γ_B) = ¼ * (6*4.00 – 3.19*4.07) = 2.75 (E coli = 27.1)

a = ¼ * (w*γ_S – c*γ_B) = ¼ * (6*4.00 – 3.46*4.07) = **2.48** (E coli = 25, no extra points off)

Thus 2.75 mol oxygen are required per mol glucose. (1 point)