• Midterms graded

• You will receive feedback and grades on HW4 by tomorrow.

• HW5 and 5-min oral presentations due in 1 week. I will let you know on Thursday whether your group will present on Tues or Thurs next week. I will also let you know which group you will review.

(Out of 93 total points)
HW5 due and 5-min presentations due in 1 week

By next week, your design project should:

• **Break new ground:** Build upon current knowledge, do not replicate previous work

• **Think big, but work small:** project goal should have a bigger picture, but it may be best to focus on a specific protein engineering technology

• **Research:** independently identify research materials and other sources

• **Communication:** It is essential that you communicate your goals and approaches clearly

• **Creativity:** Incorporate unusual sources of information and innovative ways to approach and present your project
It is essential at this point that:

• **Problem**: Your idea addresses a medical or technological problem or need

• **Main goal**: Propose to engineer a specific protein:
  – Describe desired properties of engineered protein
  – Explain how engineered protein would be used
  – Explain how engineered protein directly and logically addresses problem/need

• **Experimental approach** should describe:
  – Proposed method of protein engineering
  – Proposed number of mutated protein constructs to be investigated
  – Proposed method of protein production
  – Proposed method of assessing engineered protein function: activity assay
No bueno:

• **Problem:** cancer

• **Main goal:** we are going to engineer a therapeutic protein to kill cancer cells

• **Experimental approach:** we are going to use various protein engineering strategies and test the mutant proteins on cancer cells

It is time to be specific!!!
Protein engineering by directed evolution

- Random mutagenesis
- Oligonucleotide mutagenesis
- Recombination mutagenesis
- High-throughput screening
Evolution

- **Natural evolution**: over time, random genetic mutations occur in an organism’s genome; mutations that provide a trait that is beneficial for reproduction or survival will be preserved
  - Side note: google Darwin Awards

- **Directed evolution**: mimics natural evolution in the lab, but working at a molecular level and focusing on a specific molecular property

Similarities:
- Diversification: offspring are different from parents
- Selection: survival of the fittest
- Amplification: procreation
Directed Evolution

1. Choose a Wild Type protein
2. Make a library of mutant proteins
3. Screen the library against a target of interest and isolate best mutant
4. Is the desired property achieved?
   - Yes: End of selection
   - No: Create another library based on the best clone
5. Repeat
Benefits of directed evolution

• Does not require prediction of the types of mutations that are beneficial

• Lack of detailed knowledge in protein structure or even protein sequence homologs is compensated for by the use of a powerful selection/screening method based on the concept of survival of the fittest

• Proteins can be altered in stability, solubility, enzymatic activity, enzymatic selectivity, binding activity, etc.
Key aspect of directed evolution

Genotype-phenotype coupling:

Physical linkage between gene and protein, allowing for identification via DNA sequencing of a specific mutant protein with the desired activity

Example:
- Randomly mutate GFP gene
- Express mutant GFP proteins in *E.coli*
- Find mutants that fluoresce yellow
- Sequence mutated GFP gene from yellow colony
Directed Evolution

• Today:
  – Random mutagenesis: error-prone PCR
  – Oligonucleotide mutagenesis
  – Recombination mutagenesis: DNA shuffling

• Thursday:
  – Phage display and cell surface display
Directed evolution

Random mutagenesis

Oligonucleotide mutagenesis

Recombination mutagenesis
Random mutagenesis: error-prone PCR

Error-prone PCR:

- Create an artificial condition that is conducive of polymerase base mispairing
- Results in library of randomly mutated genes
Random mutagenesis: error-prone PCR

DNA polymerases are naturally engineered to achieve high fidelity
- Pfu polymerase: 1 error in $10^6$ to $10^7$ bases
- Taq polymerase: 1 error in $10^4$ to $10^5$ bases

Error-prone PCR: creating an artificial condition:
- Use polymerase lacking proofreading activity: Taq polymerase
- Substitute Mn$^{+2}$ for Mg$^{+2}$: 1 error in 200 to 2000 bases
- Or, use uneven concentrations of dATP, dCTP, dTTP, dGTP
- Caveat: Taq polymerase bias toward A and T mutations – may affect diversity in gene library
Random mutagenesis: error-prone PCR

Solution to Taq bias: purchase a random mutagenesis kit

Example: GeneMorphII (Agilent Technologies): a mixture of MutazymeI (bias toward G and C mutations) with Taq polymerase (bias toward A and T mutations)

→ 1 to 16 mutations per 1000 bases

Control mutation rate by varying amount of template DNA and # of PCR cycles: more PCR cycles, more mutagenesis

In general, directed evolution studies should aim for 2-7 nucleotide mutations (1-4 amino acid changes) per gene
Random mutagenesis: crunching the numbers

For a small peptide library (10 amino acids long), there are $20^{10} (= 2\times10^{11})$ possible amino acid combinations

- Same calculation applies to a protein containing 10 randomized amino acids

However, because of codon degeneracy (codon redundancy), library must be bigger than $20^{10}$ to encode a reasonable proportion of the $20^{10}$ different peptides

---

<table>
<thead>
<tr>
<th>First Position</th>
<th>Second Position</th>
<th>Amino Acid</th>
</tr>
</thead>
<tbody>
<tr>
<td>U</td>
<td>UU (phe)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UUC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UUA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UUG</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CU (leu)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CUC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CUA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CUG</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CA (leu)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CAC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CCA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CCC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CG (pro)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CGG</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CGA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CCA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ACC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AAA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AAG</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AUA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AUG (met)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>ACG</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AAC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AAA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AGA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AGU</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AGC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AGG</td>
<td></td>
</tr>
<tr>
<td></td>
<td>G (val)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GUU</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GUC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GUA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GUG</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GC (ala)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GCG</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GCC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GCA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GAG</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GAA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GAG</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GAC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GAU</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GGU</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GGC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GCA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GAA</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GAC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GGU</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GGG</td>
<td></td>
</tr>
</tbody>
</table>

How many codons for Ser?

How many codons for Met?

Note also: STOP codons
Random mutagenesis: crunching the numbers

Example: deca-Ser peptide
- Ser is encoded by 6 out of the 64 possible triplet codons
- Thus, $6^{10}$ in $64^{10}$ possible DNA sequences will encode deca-Ser

Example: deca-Met peptide
- Met is encoded by only 1 out of the 64 possible triplet codons
- Only 1 in $64^{10}$ possible DNA sequences will encode deca-Met

Take-home message: codon degeneracy should be considered when planning a peptide or protein library
(DNA libraries must be larger than calculated for amino acids)
Random mutagenesis: error-prone PCR

• Good strategy when:

  – No protein structure, little/no sequence alignments, and no good rational ideas as to where to mutate the protein

  – A high-throughput assay exists to screen high numbers of protein mutants, or a selection assay exists
Oligonucleotide mutagenesis
a.k.a. Site-directed diversification

- Create library of oligonucleotides (primers) that vary at specific region of the gene

- Results in library of genes randomly mutated in a specific region
Design of degenerate primers for oligonucleotide mutagenesis:

Order a set of primers where the wild-type nucleotides in one or more codons are replaced by mixtures of nucleotides:

- $\rightarrow$ NNN where $N = A,G,T,C$
- $\rightarrow$ NNS where $S = G,C$
- $\rightarrow$ NNK where $K = G,T$

Example:

Your sequence 5’ AAA GAC CAC AAG AAG GGG CTG 3’
Primer sequence 5’ AAA GAC CAC NNN AAG GGG CTG 3’

Caveat: may have bias toward primers with higher affinity
Oligonucleotide mutagenesis: crunching the numbers

→ NNN  where N = A,G,T,C
→ NNS  where S = G,C
→ NNK  where K = G,T

Example:

<table>
<thead>
<tr>
<th>Codons</th>
<th>AA</th>
<th>Stop</th>
<th>Prob*</th>
</tr>
</thead>
<tbody>
<tr>
<td>5’ AAA GAC CAC AAG AAG GGG CTG 3’</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5’ AAA GAC CAC NNN AAG GGG CTG 3’</td>
<td>64</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>5’ AAA GAC CAC NNC AAG GGG CTG 3’</td>
<td>16</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>5’ AAA GAC CAC NNT AAG GGG CTG 3’</td>
<td>16</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>5’ AAA GAC CAC NNS AAG GGG CTG 3’</td>
<td>32</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>5’ AAA GAC CAC NNK AAG GGG CTG 3’</td>
<td>32</td>
<td>20</td>
<td>1</td>
</tr>
</tbody>
</table>

Probability that no stop codon will occur in a peptide or protein library with 10 randomized codons. Ex: \((61/64)^{10} = 0.62\)

Note: NNS offers best codon usage overall in *E.coli*
Oligonucleotide mutagenesis: crunching the numbers

→ NNN where N = A,G,T,C
→ NNS where S = G,C
→ NNK where K = G,T

Example:

<table>
<thead>
<tr>
<th>Codons</th>
<th>AA</th>
<th>Stop</th>
<th>Prob*</th>
</tr>
</thead>
<tbody>
<tr>
<td>5′ AAA GAC CAC AAG AAG GGG CTG 3′</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5′ AAA GAC CAC NNN AAG GGG CTG 3′</td>
<td>64</td>
<td>20</td>
<td>3</td>
</tr>
<tr>
<td>5′ AAA GAC CAC NNC AAG GGG CTG 3′</td>
<td>16</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>5′ AAA GAC CAC NNT AAG GGG CTG 3′</td>
<td>16</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>5′ AAA GAC CAC NNS AAG GGG CTG 3′</td>
<td>32</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>5′ AAA GAC CAC NNK AAG GGG CTG 3′</td>
<td>32</td>
<td>20</td>
<td>1</td>
</tr>
</tbody>
</table>

Which requires the largest library?
Which requires a smaller library but still encodes all 20 amino acids?
Which one will have no issues of stop codon truncation of proteins?
How to choose between NNS or NNK?

Reminder: stop codons: TAA, TAG, TGA
Recombination mutagenesis

- Does not directly create new sequence diversity, but combines existing diversity in new ways
  - Remember: random and oligonucleotide mutagenesis creates new sequence diversity

- Recombination of two or more genes (usually homologous)
  - Degree of homology required depends on recombination method used
Recombination mutagenesis samples a larger sequence space

Random mutagenesis

Recombination mutagenesis
Homology-dependent recombination mutagenesis
1. Fragment a mixture of homologous genes by DNase digestion
   • Fragment size depends on # crossovers you want

2. Reassemble with DNA polymerase: melt, anneal, extend, repeat
   • No primers

3. PCR amplification of reassembled genes
   • With primers

W.P. Stemmer, 1994 *PNAS*
“DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution”
Comparing random mutagenesis and DNA shuffling: cephalosporinase from 4 bacteria

Cramerí et al., *Nature* 1998
Comparing random mutagenesis and DNA shuffling: cephalosporinase from 4 bacteria

Best clone: 8 segments from 3 genes

Crameri et al., *Nature* 1998
StEP (Staggered Extension Process) mutagenesis

1. Add primers to mixture of homologous genes

2. Run PCR with **VERY SHORT** extension time:
   - Melt, anneal, short extend, repeat

3. PCR amplification of reassembled genes
RACHITT mutagenesis

RACHITT: Random chimeragenesis on transient templates

1. Fragment all but one of homologous genes by DNase digestion

2. Hybridize to non-digested gene’s bottom strand

3. Digest “flaps” with nucleases and fill in gaps with DNA polymerase

4. PCR amplification

Advantage: much smaller fragments, and more crossovers, can be obtained
Problems with recombination mutagenesis

• Harder to get crossovers between genes with lower DNA sequence homology

• Crossovers occur at regions of high DNA sequence homology

• May get back original gene or limited # crossovers
Improving recombination mutagenesis when low DNA sequence homology

1. Increase DNA sequence homology with synthetic genes (without changing protein sequences)

2. Direct crossover events with oligonucleotides matching two different templates

Fragmented genes

Overlap oligonucleotide
Options in a directed evolution experiment

1. Gene library construction
2. Type of system to make protein
3. Selection/screening assay
4. Amplification of “hits”
High-throughput assays

• Screening
  – Requires inspecting each member of the library one at a time
  – Ex: screening activity in 96-well or 384-well plates with fluorescence or colorimetric measurement
  – Ex: direct screening of protein in the cell or secreted from cell with fluorescence or colorimetric measurement

• Selection
  – Relies on a property that is essential for survival
  – Excellent for libraries that are too large to screen
Overall: Directed Evolution

- Random mutagenesis creates a library of a gene that has been randomly mutated
- Oligonucleotide mutagenesis creates a library of a gene that has been randomly mutated specifically at one or more codons
- Recombination mutagenesis makes a library of chimeric genes to combine existing diversity in new ways
- Library size must be considered to cover as much sequence diversity as possible
- Codon degeneracy must be considered when calculating library size
- Your experiment is only as powerful as your detection assay