Week 9

Protein Engineering

9.1 Protein Post-Translational Modification

11/29:

- Review from last time.
  - What you need for unnatural amino acid incorporation: TGA (amber stop codon), special tRNA to both selectively bind the unnatural amino acid and TGA.
- We now study work by Wang and Schultz (2001).
- tRNA selection.
  - If we do negative selection (correct read through generates toxic barnase), nonfunctional tRNA will cause the bacteria to survive.
  - If the tRNA is only charged by E. coli synthesis, the bacteria will die.
  - If tRNA is only charged by mj syn, your bacteria will die.
  - If the tRNA can be charged by both, the bacteria will die.
  - Now for positive selection: Nonfunctional tRNA? No survival gene, so bacteria dies.
  - But this time, tRNA only charged by mj syn survives! Thus, if you do negative selection and then positive selection, you will select specifically for what you want.
  - We can also do positive selection first and then negative selection. In this case, three bacteria survive first, and then two get eliminated.
  - Positive selection seldom is the opposite of negative selection; negative does specificity, positive does ??
  - This is a very testable topic.
- An orthogonal tRNA.
  - You now have a tRNA that can only be charged by exogenous amino acid residues.
  - Orthogonal tRNA and the right gene gives you great ampicillin resistance
- An orthogonal aminoacyl-tRNA synthetase.
  - Change the specificity of *Mj* TyrRS so that it charges the selected tRNA with O-methyl-l-tyrosine.
  - The crystal structure had been determined for the homologous TyrRS from *Bacillus stearothermophilus* bound to a free tyrosine residue.
  - Five active site residues were found to be within 6.5 Å of the *para*-hydroxyl (yellow).
- Aminoacyl tRNA synthetase selection.

  - Similarly to before, it doesn't matter here if you do positive or negative selection first.
  - If you get read through for the Cm gene (chloramphenicol resistance), the bacteria survives.
  - Survivors contain aaRSs capable of charging any natural or unnatural aa onto the orthogonal tRNA.
  - We don't include the unnatural amino acid in solution for negative selection. Thus, when we have an orthogonal tRNA that only accepts O-MeTyr, these bacteria will survive because they cannot synthesize barnase. Any tRNA that incorporates an*other* amino acid will synthesize barnase and die.

- A mutant tRNA synthetase.
  - A mutant synthetase was selected byat selectively charged the *Mj* suppressor tRNA with O-methyl-L-tyrosine.
  - The substitution removes a hydrogen bond, creating a hydrophobic pocket.

- O-MeTyr as UAA incorporation into DHFR.
  - Dihydrofolate reductase (DHFR) was generated with TAG in place of the third codon and purified by metal-affinity chromatography.
  - Tandem MS of DHFR tryptic digest unambiguously shows complete incorporation of O-methyl-L-tyrosine in the third amino acid position.

- UAAs incorporated *in vivo*.
  - Reactive amino acids for further post-translational modification.
  - Caged (photo-activable) amino acids to study kinetics and mechanisms.
  - Heavy atom-labeled amino acids for structural elucidaiton.
  - Incorporated in *E. coli*, yeast, ...
  - Incorporating a bioorthogonal handle (alkyne) can be done.
  - Heavy-carbon side chain for cross-linking is possible.

- Recoding *E. coli* genome.
  - ...

- Summary.
  - Strategies to increase protein functional diversity.
    - Chemical approaches.
    - Site-drirected mutagenesis: Changing one amino acid to another.
    - Unnatural amino acid incorporation.
  - Non-sense suppression *in vitro* and *in vivo*.
  - Orthogonal tRNA and aminoacyl tRNA synthetase.
    - Directed evolution (negative and positive).
  - Applications of unnatural amino acid incorporation.

- We now start on protein post-translational modification.

- Humans are the most complex organisms on earth. But where does that complexity come from?

- Gene number and genome size.
  - We have 3 billion base pairs in our genome, containing 20k-25k protein coding genes. These constitute about 1% of our genome.

  - Do we have the highest number of genes? No! The most common species of water flea (tiny and translucent) has more than 31k genes (8000 more than us). It has so many genes due to extensive gene duplication, even though we're more complex. We also have this, but it has it more.
  - *C. elegans* is a model organism that is often used, especially in aging studies since it contains all of the human aging genes. It's life cycle is just much shorter and therefore easier to study. It has a similar amount of protein-coding genes to humans!
  - Do we have the largest genome by base pairs? The Japanese flower *Paris japonica* has a 149 billion base pair genome, 50 times the size of our haploid genome.
  - So why are we so complex?
  - We are complex due to meticulous regulation of transcription and translation (the other 99% that's not non-coding is largely regulatory; we don't fully understand it, but we know it's not just junk).
  - Proteome complexity also plays a role.

- **Proteome**: The entire set of proteins that is, or can be, expressed by a genome, cell, tissue, or organism at a certain time.
- **Proteomics**: The study of the proteome, such as proteome profiling.
- There's genome, proteome, **transcriptome**, **kinome** (all kinases), **methylome** (all methylated DNA), etc.
- **Kinome**: The complete set of protein kinases encoded in its genome.
- Post-translational modifications augment proteome complexity.
  - Chemical changes that happen after translation.
  - Lactylation, other example??
  - Common types: Phosphorylation, hydroxylation, glycosylation (N-glycosylation from Dr. Krishnan; super important for cell recognition; Dr. Tang use to have a lecture on it), Lysine acetylation, lysine ubiquitinylation (adding a 76-77 aa peptide instead of a small functional group; usually precedes degradation).
  - You don't have to memorize these; if tested, we will be given a chemical structure.
- Protein phosphorylation.
  - Catalyzed by kinases.
  - Within the kinase active site, you have a base that deprotonates the hydroxyl group, leading the O^- nucleophile to attack the γ phosphate in an ATP via nucleophilic acyl substitution.
  - Protein phosphatase removes phosphates from phosphorylated protein residues, resulting in...
- Functions of protein phosphorylation.
  - 1/3-2/3 of the proteome in eukaryotes can be phosphorylated.
  - Not all of this leads to activity (as far as we know at this point).
  - Phosphorylation alters...
- The human kinome comprises 518 kinases.
  - Tyrosine kinases have their own corner of the evolutionary phylogenetic tree.
  - There are 48 FDA approved kinase inhibitors; more than half are for tyrosine kinases.
  - What's the difficulty? Do we want to inhibit one tyrosine kinase or many? Probably just one, belonging to one protien. Thus, *specificity* is the largest challenge of developing novel kinase inhibitors.

  - Gleevic (or Imatinib) is a drug for Leukemia. Once the...
- MAP kinase pathways: An example of kinase cascade complexity.
  - MAPKKKK lol.
  - One protein can be phosphorylated by many kinases.
- Mass spectrometry-based **phosphoproteomics**.
  - ...
  - There are about 200 phosphorylation sites in the human genome.
  - We don't know the functions of most phosphorylation sites.
- **Phosphoproteomics**: A branch of proteomics that characterizes phosphorylated proteins.
- How can cellular substrates of specific kinases be identified?
  - *In vitro* peptide microarray screens.
  - You take a chip, divide it into 1000-2000 wells.
  - Each well encodes a specific peptide sequence.
  - Then you flow on the kinase you want to study and supply radioactive ^{32}P. Then you can take a radiograph of the chip.
  - No longer widely used because it's not super biologically relevant.
- Tyrosine kinases.
  - Shokat's bump and hole strategy for kinase substrate identification.
  - His lab developed one of the first covalent drugs for ?? inhibition (cancer related).
  - Make ATP larger so that...
- Interaction between mutated kinase and modified ATP.
  - ...
- Kinase/ATP analog engineering allows substrate tagging in cell lysates but not in live cells.
  - ...
- Use chemical tagging...

9.2 Bioorthogonal Chemistry

12/1:
- Nobel prize in Chemistry (2022).
- Outline.
  - Bioorthogonal chemistry: What and why?
  - Bioorthogonal reactions.
  - Applications of bioorthogonal chemistry.
- **Bioorthogonality**: Chemistry inert to the conditions in the particular physiological condition.
  - Non-native reactants and selective reactions under physiological conditions.
- Understanding biology through chemistry.
  - **Chemical biology** defined.

  - Key to the field is selectivity.
    - Chemical biologists hope to develop molecules with perfectly selective biological function.
    - In practice, though, we start with as much as we can get on a first attempt and refine from there.
  - The ability to make chemical modifications that enable direct detection of, or interaction with, biomolecules in their native cellular environments is at the heart of chemical biology.

- **Chemical biology**: The creation of nonbiological molecules that exert an effect on, or reveal new information about, biological systems.

- Advantages of bioorthogonal chemistry.
  - Applicable to all biomolecules (in theory).
  - Small size (non-perturbing; better access to intracellular and extravascular compartments).
  - Bioorthogonality = selectivity (highly selective, low background labeling *in vivo* — in whole organisms!).
  - Versatile and divergent (two-step labeling enables various functionalizations of the same reporter group).

- Bioorthogonal reactions.

- Requirements of bioorthogonal chemistry.
  - Definition of a **bioorthogonal reaction**.

- **Bioorthogonal reaction**: A chemical reaction that satisfies the following conditions.
  1. The reaction must be chemically selective and compatible with an aqueous environment.
  2. Reaction yields a stable covalent linkage without toxic byproducts.
  3. Reactnats must be kinetically, thermodynamically, and metabolically stable, and nontoxic prior to reaction.

- Important reactions:
  - Ketone condensation, e.g., with a labeled hydroxime.
  - Staudinger ligation, with an aza-ylide (nitrogen and phosphorous interacting).
  - CuAAC (see Figure 2.3).
  - SPAAC (same idea as the above, except the alkyne is in a ring).
  - Tetrazine-BCN, with a 4x N ring that loses N_2 and bonds with an cycloalkene instead.
  - Different rate constants for various reactions shown.

- Ketone condensation.
  - Reactions of aldehydes/ketones with amine nucleophiles (usually hydrazine/alkoxyamine).
  - Usually requires acidic pH, slow kinetics, mM concentration of reagent, competition from endogenous or naturally occuring aldehydes and ketones.
  - Aniline accelerates the reaction ¿ 40-fold (¿400-fold at pH = 4.5).
  - Pictet-Spengler ligation of the aldehyde with an alkoxyamine derivative is best.

- Staudinger ligation.
  - Azide reacts with $RPPh_2$ under mild conditions.
  - Internal electrophilic trap forms amide linkage.
  - Phosphines are relatively unreactive toward biological functional groups.

  - Reaction is relatively slow.
- Cu(I)-catalyzed azid-alkyne cycloaddition (CuAAC).
  - Azide (1,3-dipole) can undergo reactions iwth activated alkynes.
  - Forms triazole products but at physiological conditions.
  - Fast, but high cellular toxicity.
- Strain-promoted $[3 + 2]$ cycloaddition.
  - Strain-promoted azide-alkyne cycloaddition (SPAAC).
  - Catalyst free $[3 + 2]$ (toxic Cu(I) not needed).
  - Can be performed on the surface of living cells.
  - Increase reaction rate with the addition of an EWG on cyclooctyne.
- Nitrone cyclooctyne reactions: $[3 + 2]$ cycloadditions.
  - Strain-promoted alkyne-nitrone cycloadditions (SPANC).
  - Uses more reactive 1,3-dipole nitrous in place of azides.
  - Rate constants up to $60\,M^{-1}\,s^{-1}$ (60-fold faster than SPAAC).
  - Faster rates means lower reagent concentrations can be used.
  - Cyclic nitrones are more stable than acyclic counterparts.
- Strained-promoted alkene/alkyne-tetrazine reactions (SPATL).
  - 1,2,4,5-tetrazines are reacted with electron rich dienophiles (alkenes).
  - 3,6-diaryl-s-tetrazines were found to be stable in water.
  - Run in cell media and cell lysate with ¿80% yield.
  - Successfully used to label proteins in vitro and in cells.
- Applications of bioorthogonal chemistry.
- Introducing ketones with biotin ligase.
  - ...
- Fluorescent labeling of proteins on ketones.
  - Cell surface proteins can be tagged with the AP at their N- and C-termini.
  - Then attach a ketone moiety, then attach a fluorescent hydrazide.
  - Labeling is efficient and specific when performed on purified proteins, total cell lysates, and intact cells.
- Trafficking of labeled EGFR.
  - ...
- Zebrafish embryogenesis.
  - Answers questions like, “how can you tell the relative levels and types of cell surface glycosylation during the course of embryogenesis and development?”
- Using alkyne bioorthogonality to study glycosylation.
  - Feed cells N_3-sugar and see where reactions take place in real time.
  - Spatio-temporal analysis of glycosylation during embryogenesis and development.
  - Different fluorophore reagents and different bioorthogonal chemistries let you see the full picture.

9.3 Final Review Sheet

2/2/24:

- Watson-Franklin-Crick vs. Wobble interactions.
  - The reason some codons are the same is in case normal Watson-Franklin-Crick interactions get supplanted by Wobble interactions, we don't want the amino acid paired to change.
- At physiological pH, only phosphates are charged.
- We can get tautomerization among nucleoside bases; think keto-enol.
- Puckomers: Ribose conformers.
- Why phosphate is cool: Multivalent, cannot cross biological barriers, kinetically stable to hydrolysis, thermodynamically unstable, kinetically unstable with catalyst.
- A-, B-, and Z-DNA.
- G-quadruplexes.
- Tagging proteins with GFP, and RNA with the spinach aptamer.
- DNA binding proteins.
  - Interact w/ DNA major + minor groves.
  - Leucuine zipper and zinc finger; precursors to CRISPR.
  - Intercalators: Like the toxic molecule ethidium bromide.
    - Design safer ones by making bigger molecules; ones that still intercalate DNA but don't penetrate the skin as easily.
- DNA replicase.
  - One Mg^{2+} stabilizes the dNTP's two extra phosphate groups; the other stabilizes the acyl substitution intermediate.
  - Repairs the strand by sliding back to check previous bases before moving on.
- The causes of mutations.
  - Natural mismatching and tautomerization.
  - Deamination of exocyclic amines.
  - Depurination and cleavage.
- Four strategies of DNA repair.
  - Direct reversal/repair: Enzymes catalyze the reverse reaction of whatever nefarious transformation (e.g., deamination) took place.
  - Base excision repair: Take out a base, put in a new one.
  - Nucleotide excision repair: Take out a string of nucleotides, synthesize a new one.
  - Mismatch repair: Take out a mismatch on the *unmethylated* strand, resynthesize.
- Bioorthogonal chemistry and the click reaction (azide plus alkyne).
- Promoter regions (for RNA synthesis).
- Nucleic acids may have catalyzed reactions in early forms of life.
- *Tetrahymena* Catalytic RNA.
- Mechanism of the ribosome: A2486 catalyzes protein formation via proton transfer.

- Bisulfite chemistry to detect methylated cystine.
  - Bisulfite plus heat converts cystine to uracil, so sequence once on its own and once after bisulfite chemistry and look for which base pairs don't change.
- Almost all amino acids are in their chiral L-form.
  - We can build D-proteins, though.
  - These cannot be degraded by natural protease and hence are much more stable.
- Native chemical ligation: Connecting two peptides with an amide bond.
- Know a bit about how to draw nucleosides, and the amino acids.
  - Achiral.
    - Glycine: Flexible. GGS linkers are nice.
  - Hydrophobic.
    - Ala/A: Simple; good to mutate things to to determine importance. Often works as an inert filler.
    - V,L,I: Use these for bulk, e.g., to decrease the size of active sites and make pockets smaller.
    - M: Good start codon. Part of the cofactor SAM, a methyl donor.
    - Proline: Inflexible.
    - Phe/F, Tyr/Y, Trp/W aren't too interesting.
  - Charged.
    - Acids deprotonated and bases protonated at physiological pH.
    - Asp/D, Glu/E, Lys/K, and Arg/R aren't too interesting.
    - His/H is a great base/acid for a proton shuffle.
  - Polar.
    - Ser/S can be phosphorylated (just like Y). Often a nucleophile in protein active sites.
    - Thr/T, unimportant.
    - Cys/C forms disulfide bridges, unlike M.
    - Asn/N is often a metal coordinate (think Mg^{2+} ions in DNA polymerase!).
    - Gln/Q is similar to N.
- Protein structure in 4 levels: Primary, secondary, tertiary, and quaternary.
- Serine protease.
  - Asp/D, His/H, and Ser/S work together to cleave a protein.
- Structural biology stuff.
  - Lots of XRD.
  - LCLS.
  - NMR (for in-situ; higher dimensional).
  - CryoEM (for big stuff).
  - AlphaFold 2.
- PCR and thermal cycling with primer.
- DNA sequencing.
  - Maxam-Gilbert.

    - This is the one with variable cleavage reagents and separation in a gel.
  - Sanger.
    - Spike a bit of ddNTP into cloning bath to stop cloning at a certain point.
    - Either do sequential (ddATP, ddGTP, ...) or parallel (ddNTPs plus fluorophore).
  - Pyro (454).
    - Create a bunch of copies, immobilized via biotin and streptavidin beads.
    - Add a specific type of dNTP to be incorporated, releasing a PPi, which gets converted to ATP and a flash of light that can be detected by two consecutive enzymes.
    - Get rid of excess dNTP and repeat.
    - You can do this in a well plate to run many tests in parallel.
    - Important for neanderthal genome.
  - Illumina.
    - Currently the most important.
    - Bridging boi.
    - Read each strand individually by flashes of light, though there will be multiple copies in each cluster, so multiple photons will be released.
    - We then sort through the data to align all pairs.
  - SMRT.
    - Fix DNA polymerase in a zero-mode waveguide (tiny pore), run a sequence through it recording flashes as each dNTP + fluorophore is added.
    - DNA polymerase is slowed down to a recordable speed by attaching a protecting group to each dNTP's $3'$ site.
  - Nanopore.
    - Still in development.
    - Electrical rather than chemical.
    - A single strand passes through a pore; each dNTP blocks ion flow through the pore to a different, unique, detectable extent.
    - Allows for direct sequencing of methylated bases, too!

- The fluid mosaic model of the plasma membrane is *very wrong*.
  - 500-2000 different kinds of lipid molecules.
    - Different alkyl chain lengths and degrees of unsaturation.
    - Different head groups, too.
  - 17-23% cholesterol.
  - Asymmetric; different things face in vs. out.
  - There are many different ways a protein can stick into the plasma membrane.
    - Single-pass (via an amphipathic helix), multi-pass, lipid anchoring, GPI anchoring, β-barrel proteins.
    - Predict transmembrane domains with the hydropathy index.
  - Know when a cell has died via annexin staining; PS phospholipids are constantly flipped inward while the cell is alive, but when it dies, it can no longer do this, signalling to immune cells that it is dead.
  - Extracellular proteins can bend the plasma membrane.
- There are various kinds of transporters and transportation.
- Chaperones bring misfolded proteins to the right place and allow them to fold correctly.

- Remember topologically equivalent compartments.
- Isolating organelles by progressively fast centrifugation.
- N-terminus localization sequence for proteins.
- Nuclear pores.
  - Membrane ring proteins to bend nuclear membrane around pores.
  - Cytosolic fibrils and nuclear basket.
  - Nuclear import adapter proteins.
  - Ran-GTP/Ran-GAP in and Ran-GDP/Ran-GEF out.
- Mitochondrial transport.
  - TOM: Outer to interluminal.
  - TIM: Interluminal to inner.
    - Locks together with TOM to do outer to inner directly.
    - Translocation sequences get cleaved once inside inner matrix.
    - Inner membrane protein? If TIM encounters a stop transfer sequence.
      - ➢ Multipass? Stop transfer sequence in the middle.
  - SAM: Insertion into outer membrane.
    - Takes proteins brought to interluminal space by TOM.
  - OXA: Interluminal to inner.
    - Can do in inner membrane if it encounters a stop transfer sequence in a protein in the matrix.
  - To get into the interluminal space, either insert into inner membrane and cleave w/ signal peptidase *or* get in through TOM and fix there with Mia40 (forming disulfide bridges between different proteins that prevent it moving further into the matrix).
- Getting proteins into ER lumen.
  - SRP grabs ribosome with growing peptide with signal sequence, brings it to ER membrane receptor, portein grows into ER lumen and is left there.
- Getting proteins into the ER membrane.
  - Snare proteins grab C-terminus localization sequences. Get1, Get2, Get3 over to ER membrane and insertion.
- GPI anchors.
  - Lipids to phosphate to many sugars to protein.
- Golgi.
  - Stacks are cisterni.
  - Golgi cisternae kind of grow up. Either the **maturation model** or the **vesicle transport model**.
- RNA interference.
  - Degradation of mRNA triggered by homologous dsRNA.
  - Something with double- vs. single-stranded RNA.
- Unnnatural amino acid incorporation.
  - CRISPR-Cas9, guide RNA, selective cleaving and then cellular repair integrates what you want.
  - Unnatural amino acid incorporation via bioorthogonal tRNA to allow you to control protein structure.