# 5.08 Fundamentals of Chemical Biology

Notes from Spring 2020.

Last Updated: May 16, 2020.

# Contents

# 1 Introduction

These notes were typeset live in 5.08 lecture. Special thanks to Andrew Lin for providing his LaTeX formatting template.

Today's lecture is being taught by Prof. Imperiali, and we'll be going through an overview of the course and a bit about nucleic acids. We'll first talk about some guiding questions of chemical biology:

**Question 1.** *What is Chemical Biology?? Can you define it? What types of studies exlemplify chemical biology? What are the big frontiers?*

There are many right answers to these questions, and each chemical biologist has their own answer. Chemical Biology is about the genesis of tools and approaches (mostly from chemistry, but also physics and engineering) that allow us to study biological systems. Essentially, chemical biology is an interdisciplinary tool that gives us more insight into biology. We utilize chemical principles to manipulate systems to investigate the underlying biology (**systems biology**) or to create new functions or materials (**synthetic biology**). Biochemistry investigates the chemistry that underlies biological systems; chemical biology attempts to understand biological systems themselves.

**Note 2**

For a reference, see "Voices of Chemical Biology" *Nat. Chem. Biol.* 2015, 11, 378.

**Example 3**

Not only is chemical biology a broad topic, but there are also broad applications, including:

- Site directed mutagenesis (Molecular Biology)
- Sequence analysis (Bioinformatics)
- Structure analysis (Biophysics)
- Intermolecular interactions (Biophysics)
- Yeast screening (Molecular Biology)
- Chemical Synthesis (Organic Chemistry)

Now we're going to move onto course logistics. The course is designed to encompass the above main tools. We'll cover the 6 major topics as follows:

- Nucleic Acids
- Proteins and Enzymes
- Biophysical Methods
- Manipulating the Protein Architecture
- Post-Translation Modifications
- Frotiers of Chemical Biology

Prof. Imperiali also emphasizes the importance of communication between chemists and biologists, with a brief historical interlude:

**Fact 4**

The first surge of chemical biology was in the 1950s, but there were fundamental disagreements between biologists and chemists. Biologists often saw chemists as annoyances; no dialogue resulting in very few problems being solved. However, there is now resurgence of chemical biology all around. "Biologists have the problems. Chemists have the tools."

An example of the failure of communication was on a bacteria in Mono Lake:

**Example 5**

This bacteria was discovered to grow on arsenic rather than phosphorous. Biologists proposed that arsenic completely replaced phosphorous in proteins and nucleic acids; chemists thought that this didn't make any sense. No spectrophotometric measurements done; later found that transporters within the cell still heavily favour phosphate. A 10000x increase in arsenic only results in 3x more arsenic within DNA (still extremely small). Chemical studies showed that half life of phosphate = $3 \cdot 10^7$ years, that of arsenate = 0.06 s. So, talk to chemists!

Most of the grading is from homeworks (60%); the last homework is a proposal. Late policy - 90% credit for one day late, 80% for two days late, no credit for later. There will be 5 take-home quizzes (mostly of 'vocab', 15%). If you can't trust yourself, take it with a group - "Misery loves company." There is also a final Presentation (25%) at the end of the semester, teams of 2. Graduaute students have presentations through the course.

# 2 Lecture 1: February 3, 2020

We're now going to talk a little about nucleic acids. As we know from previous courses, we have the central dogma of information transfer: DNA to RNA to protein. We can also make new variations on the central dogma, namely post-translational and post-transcriptional modification (both pre-mRNA splicing and protein splicing, etc) Each step can be modified (for example, introducing unnatural amino-acids, etc).

The structure of nucleic acids were originally covered in Franklin's X-ray diffraction experiment and Watson/Crick's double helix proposal.

**Note 6**

It's important to not learn wrong structures - no carbons with 5 bonds, etc.

"I'm too lazy to talk about nucleic acid synthesis today, so we'll do it on Wednesday."

**Question 7.** *How did DNA come to make DNA?*

It's like the chicken and egg problem. You need DNA to make DNA! And if its DNA polymerase, you need a primer and a DNA to make DNA. However, the state of the art DNA synthesis is not dependent on either primer nor DNA!

# 3 Lecture 2: February 5, 2020

Today's lecture is being taught by Prof. Raines. We'll continue with our previous discussion of nucleic acids.

The central dogma, involving the transcription of DNA to RNA then translation to protein, gives both a way where we can understand the process and how it can be modified to suit our needs.

**Fact 8**

In the 1900s, Linus Pauling and Chargaff met on a ship, and Chargaff mentioned his work about the measurements of purine and pyrimidine equivalences. Pauling ignored this due to somewhat of a personal vendetta and his model of DNA, published in Nature in 1953 and PNAS, had a triplex DNA structure with phosphate groups on the outside (assuming that DNA was an ionic crystal and that phosphate groups were protonated, impossible). This was revealed by Crelin Pauling (the son of Linus Pauling) in 1995.

Watson and Crick ended up publishing their correct structure two months later, and all that was left for was experimental verification.

> "What I cannot create I do not understand" -Feynmann.

The first attempt at synthesis of nucleic acids was with a DCC-mediated phosphate coupling. From Prof. Raines' own experience, it's really ugly chemistry, the yields are low, and there are many side products.

Now, solid phase methods are much more common in synthesis of biopolymers. The first example was merrifields' solid phase synthesis of peptides (Nobel Prize 1984), and has been extended to Carruthers' solid phase nucleic acid synthesis.

**Question 9.** *How is DNA integrity preserved? After all, the $t_{1/2}$ of phosphodiester bonds is 30 million years (Ref: Wolfender 2006, 103, 4052), so why can't we get DNA of dinosours?*

**Fact 10**

The amines of the nucleotide bases can become hydrolzed, for example, cytosine can be hydrolyzed to uracil. This happens about 100 times per human cell per day, but cells have enzymatic correcting methods. There are also other enzymes, such as adenosine deaminase, which convert adenosine to inosine, which allows us to manipulate the genome. David Liu (Harvard) is a pioneer of such methods.

Phosphates are also an integral part of DNA, especially owing to the long half life and stability of phosphodiester bonds. A more thorough discussion is given in "Why Nature Chose Phosphates" by Westheimer.

**Fact 11**

There is quite a misconception of the phosphate anion. The common depiction has the phosphate as an unsymmetric molecule with a double bond, but a more accurate depiction is given by a phosphate with $T_d$ symmetry and negative partial charges on all oxygens which protect the phosphorous from nucleic attack. "You'd be hard-pressed to find a biochemistry text that depicts the phosphate in this way" (Anslyn does)!

Now we'll get to the meat of Carruthers' DNA Synthesis. The basis of Carruthers' synthesis is to start with a protected nucleotide, with the 5'-hydroxyl protected by a dimethoxytrityl (DMT) group and the nucleobase being protected by either a benzoyl or N-2 isobutryoyl group. On the 3'-hydroxyl, there is a phosphoramidite group (phosphorous in +3 state), connected with diisopropyl amine and an O-ethylcyanide.

The DMT protecting groups from the ribose can be removed from with acid, and the product results in an orange color, allowing us to monitor the course of the reaction. The protecting groups on the nucleobases can be removed from base-labile. To form the phosphodiester linkage, a 1H-tetrazole first substitutes the diisopropyl amine, and then the 5'-hydroxyl of the new base attacks the phosphorous and makes the tetrazole a leaving group. Oxidation with iodine oxidizes the phosphorous to +5, and basic conditions remove the ethylcyano group.

This doesn't really work with RNA, however, because the 2'-hydroxy group will cyclize with the phosphate group in basic conditions. So, we need to protect the 2'-hydroxy group with a group that is orthogonal to both acidic and basic conditions, namely a TBDMS group. Other proposed suggestions were benzyl groups (hydrogenlysis) or with an Alloc group, but the logistic difficulties of hydrogenlysis or the expensiveness of palladium removal of Alloc make the silyl groups the best.

The danger of this solid-phase method is if one nucleotide is not attached accidentally, then it is super hard to seperate between a 99-length and 100-length polymer. So, you cap the hydroxyl group after each step with something like acetic anhydride to stop that chain growth.

**Note 12**

There is an online website that explains this procedure in detail: `https://www.atdbio.com/content/17/Solid-phase-oligonucleotide-synthesis`

Previously, we would have machines in our own labs to make whatever oligonucleotides we wanted. Now, there are automated DNA synthesizers that make these and get them delivered within 48 hours.

There are also nucleotides that don't form hydrogen bonds, but they are also forming double-stranded helixes (See: Eric Kool; Hirao: *Curr. Opin. Chem. Biol.* **2006**, 10, 622). This implies that hydrogen bonds do not play as crucial of a role as previously thought in maintaining stability. These new nucleobases form more stable duplexes, but are not amendable to DNA-polymerase mediated polymerization. You can get extremely versatile structures from modification of the ribose (xeno-nucleic acids) that actually get polymerized by DNA polymerase! (Ref: Hollinger, *Science* **2012**, *336*, 341). In fact, Eschenmoser showed that 6-member sugars can even be incorporated into DNA, and these are also found to be more stable than 5-member rings!

**Question 13.** *Why did nature use 5-member rings instead then?*

The answer is unknown, but it is speculated that nature did so so that DNA could be taken apart as well for cellular purposes.

**Note 14**

A non-natural base-pair is essential for the propagation of the plasmid within the E. Coli. (Romesberg, Nature **2014**, *509*, 385). In fact, this non-natural base-pair was necessary for the survival of the plasmid, and this was the first example of a non-canonical base pair in DNA. There is now a lot of research in non-canonical nucleic acids.

We'll take a quick interlude to the "Most beautiful experiment in Biology," which demonstrated the semiconservative nature of DNA replication. First, a double-stranded $^{15}N$ DNA was made, and was allowed to grow in $^{14}N$ medium. This resulted in two double stranded DNAs, each half $^{14}N$ and $^{15}N$. Another round of replication revealed that you got pure $^{14}N$ DNA only, which shows the semiconservative nature of DNA replication. This was ultimately measured with gravitational separation (since $^{15}N$ is heavier than $^{14}N$). In fact, this experiment was actually done earlier, but with iodine replacing the methyl group in thymine as the tracker instead by Meselson in his PhD thesis, under Linus Pauling. This was however met with criticism since it was thought that iodine could perturb DNA replication, even though it was chemically similar.

For actual DNA manipulation, we can use **restriction enzymes**, which were discovered in the 1960s as a defense mechanism in E. coli. They are proteins that recognize palindromic DNA sequences and destroy them. The E. coli protects its own DNA by methylating their own sequences. Now, restriction enzymes are becoming phased out as well, and have mostly been replaced by the Polymerase Chain Reaction (PCR) and the Gibson Assembly.

**Note 15**

References:

**PCR:** Khorana: J. Mol. Biol. **1971**, *56*, 341; Mullis: US Patent 4683195 (filed 1986); Taq polymerase - Trela: *J. Bacteriol.* **1976**, *127(3)*, 1550.

**Gibson Assembly:** Nature Methods, **2009**, *6*, 343.

# 4 Lecture 3: February 10, 2020

Today's lecture is given by Prof. Raines. We'll be discussing some more facets of DNA and DNA manipulation.

**Fact 16**

In 1977, the rate of DNA sequencing was about 1 base pair per month. Now, with Sanger sequencing and Gilbert sequencing we can sequence the human genome ($3.2 \cdot 10^9$ base pairs) in a few hours. The cost of sequencing originally followed a decreasing log-linear distribution (similar to Moore's law), and with the advancement of next generation of DNA sequencing after 1977, has decreased even further than expected.

The method of DNA sequencing developed in 1977 was known as **Sanger sequencing**, also known as chain-termination sequencing. The way this process worked was that there were four tubes, each containing the following:

- DNA to be sequenced
- Oligonucleotide primer
- All four dNTPs
- One type of $^{32}$P-labeled ddNTP
- DNA polymerase

Since the ddNTPs do not have a 3'-hydroxy group to continue polymerization, they essentially act as caps for the growing DNA molecules. They are present in much smaller concentrations than the dNTPs, and each uniquely caps one base pair. Probabilistically, this means that each DNA fragment may get terminated at a certain nucleotide. Repetition with the other ddNTPs allows for complete sequencing of the DNA. After completing the reaction, the DNA from all four tubes were subject to gel electrophoresis and exposed to an X-ray film. This gel electrophoresis has resolution of individual DNA molecules. This process has mostly stayed the same through the decades, but has had some innovations, such as replacing radioactive nucleotides with fluorescent ddNTPs. Since each of the ddNTPs are uniquely labelled with a different color we can now use capillary electrophoresis, which can be scanned by computer.

**Remark 17.** *Prof. Raines presents a gel that he himself ran a long time ago. He calls it 'antique,' and mentions that the 'antique' instruments he used may be found somewhere on campus.*

Now, even Sanger sequencing has been replaced, with **Next-Generation Illumina Sequencing**. This works by having a chip with many small oligonucleotides attached to it, and then DNA

fragments (about 100 nucleotides) recognize the oligonucleotides and the oligonucleotides copy the DNA fragments. This amplifies the DNA and makes about 1000 copies of a fragment, and a computer tracks the synthesis of DNA and puts it back together.

`https://www.youtube.com/watch?v=fCd6B5HRaZ8/` provides a good introduction to this process.

We'll now talk a bit about DNA's 3D structure. DNA has both a **major groove** and a **minor groove**, and proteins largely recognize DNA by the major groove, which an $\alpha$-helix of a protein fits well into. Arginine is very good at forming hydrogen-bonds with Guanosine, and asparagine and glutamine are well-equipped to recognize Adenosine. Another method of a protein recognizing DNA is by the **coiled-coil**, or leucine zipper, which is when two $\alpha$-helicies bound together. It binds as a dimer and has interactions that demonstrate cooperativity. This motif is found in many proteins, not just those that bind to DNA.

**Note 18**

The name of leucine zipper comes from the fact that the area of binding of the two alpha helicies are often leucine molecules, and it is hypothesized that the isopropyl groups on the leucine side-chains can interlock. However, this is probably not true, and now this term is not used often.

Another motif, often only found in transcription factors, is the **helix-turn-helix**. It is a short $\alpha$-helix lies in the major groove and is stabilized by another short $\alpha$-helix, separated by a turn in the DNA.

A third motif is known as the **zinc-finger**. It commonly consists of two cysteine and two histidine residues (but there are more possibilities!), which all coordinate to a $Zn^{2+}$ ion which organizes the structure of the protein, often making an $\alpha$-helix more stable, and it is this helix that interacts with DNA. This is one of the more modular and stable motifs. The zinc is essentially redox inert and acts as a good coordinating agent.

**Note 19**

The major reference for these protein-DNA interactions is a review, Wolberger, *Mol. Cell.* **2001**, 8, 937.

Small molecules can also interact and bind with DNA. Some molecules, such as Ethidium Bromide (DNA stain) or Doxorubicin (cancer drug), can intercalate with DNA, through interactions involving both H-bonding and pi-stacking. Other dyes, such as Hoeschst 33258, can bind instead in the minor groove through H-bonding (though non-specific), and can change color when bound to the DNA.

Other natural products, such as distamycin, can also bind in the minor groove, and can also show specificity through H-bonding.

In low concentrations, distamycin bonds in a 1:1 ratio, interacting with both DNA strands, but at higher concentrations, a 1:2 complex between distamycin and DNA is formed, with each molecule of distamycin binding only with one strand of DNA, mostly with A/T base pairs. Now, if these two molecules of distamycin are connected by a flexible link, and the pyrrole ring is replaced with imidazole, then this modified distamycin (now known as **polyamides**) has extremely high specitifity for DNA.

**Note 20**

Ref: Dervan, *Bioorg. Med. Chem.*, **2001**, 9(9), 2215.

Now that we've developed models for interactions between molecules and DNA, we can see how these molecules can actually result in the selective cleavage and modification of DNA. In order of specificity, we have meganucleases, zinc-finger nucleases, TALEN, and CRISPR-Cas. Meganucleases function similarly to a restriction enzyme (about 20-40 bp of recognition), while zinc-finger and TALEN uses a **fok-nuclease**, a protein sidechain, to recognize DNA, to achieve about 1-3bp of specificity. CRISPR-Cas uses modern technologies to actually read the DNA itself, and uses base-pairing to uniquely specify base pairs.

**Note 21**

CRISPR Reference Videos:

`mediaspace.wisc.edu/media/movie_CRISPR_1/0_h6dlrzdw/`

`mediaspace.wisc.edu/media/movie_CRISPR_2/0_zpjluc9m/`

The origin of CRISPR/Cas9 was originally from bacteria, which used it to recognize viral DNA sequences through a sequence of about 20 base pairs that paired with unzipped DNA. But natural selection means that these viruses have now created Cas9 inhibitors that essentially disables the Cas9 recognition machinery. CRISPR is good at detecting and cleaving genomes, but can also cause side effects, so now inhibitors may be used to essentially shut CRISPR down after a few seconds.

Now, there is quite a bit of research on CRISPR techniques, and genomes can be modified to create significantly different phenotypes (for example, *Nature* **2015**, *523*, 13.) Few human genetic diseases are actually curable with CRISPR alone, since it only allows for 'cutting' of the DNA sequence.

**Note 22**

There are also CRISPR-kits for sale, that allows essentially anyone to modify DNA with CRISPR techniques.

How do we edit DNA instead of just cutting it? Recall from previously that cytosine undergoes spontaneous hydrolysis to uracil. But there is also another enzyme **cytidine deaminase** that catalyses this process!

**Note 23**

Uracil will code like Thymine when in DNA, and Inosine (product of adenosine deaminase) will code as a Guanine residue.

When we do this process to a C - G base pair, the C becomes replaced by a T. CRISPR/Cas9 recognizes the complementary strand and cleaves it, and then the cell's own DNA repair machinery recognizes that this is not stable, and it replaces the G with an A, thus essentially changing C - G to T - A. A similar method with adenosine deaminase will change an A - T base pair to a C - G!

# 5 Lecture 4: February 12, 2020

Today's lecture is again given by Prof. Raines, and will continue talking about modifying DNA. Last time we talked about spontaneous reactions with exocyclic amino groups, but we can also cause **transversions**, which are when purines and pyrimidines are interchanged. For example, when guanine gets oxidized at the 8-position, the syn variant now pairs with adenonine instead of cytosine.

**Note 24**

Note that base editors can only repair transitions, but not transversions.

To further modify DNA, we have an enzyme **reverse transcriptase** which can transcribe RNA back into DNA. However, it has no proof-reading ability, which means that it has a much higher error rate. The nice thing about reverse transcriptase is that it can actually be paired with CRISPR-Cas9's guide RNA to synthesize entirely new DNA (known as a prime editor)!

**Remark 25.** *Though there are still quite a few issues (Ref: Liu, Nature,* **2019***, 576, 149), this work has essentially made it possible to modify DNA directly into nearly any sequence we want.*

**ChIP Seq (Chromatin Immunoprecipitation Sequencing)** can additionally be used to identify DNA sequences that interact with specific proteins, such as histones, transcription factors, RNA polymerase, DNA polymerase, and DNA-repair enzymes. The process first works by using formaldehyde to cross-link bound proteins to DNA, then sonication to sheer the DNA, then precipitating the chromatin with an antibody that has specificity for the protein of interest. Removal of the cross-link and protein, results in the DNA fragments that the protein binds to.

Overall, in the modification of DNA, and biological molecules as a whole, a new way of thinking about chemical biology is thinking about molecules in a modular fashion, where essentially every part of a biological molecule can be thought of as a module that can be interacted with. These modules also offer possibilities for replacemenet by either other biological molecules or even synthetic compounds.

**Example 26**

One example of this is in human hormones, which cause transcription. Normally, the vitamin D-receptor binds to vitamin D (obviously!). However, sometimes there is a mutation that chances Arg274 to Leu, which reduces the necessary hydrogen bonding interaction to vitamin D and causes Rickets disease. Instead, by replacing vitamin D with a synthetic agonist, then activity can again be increased.

Now we're going to shift gears a little bit and talk instead about RNA. RNA instead exists as the **A-form**, but we don't have the technology to recognize RNA sequences (as compared to Dervan's DNA technology). The major difference between RNA and DNA's 3D structure is not the actual prescence of the hydroxy group, but rather the puckering caused by the furanose ring. The difference is potentially because of gauche effects, caused by donation of lone pairs into antibonding orbitals that promote the endo isomer. This is relevant as the C3' endo isomer is preferred, but the C2' endo isomer is preferred when the C2' hydroxy group is removed (for example, with ribonucleotide reductase). Another potential factor is the effect of more hydration in RNA, which also lowers overall system energy. The puckering of the ring allows for the hydroxy groups to be more accessible to the solvent.

**Note 27**

In DNA, a repair mechanism will methylate uracil bases to make them thymines, under the action of thymidylate synthase. This enzyme is inhibitated by 5-fluorouracil. It is known as a **mechanism-based inactivator**, since the mechanism for the enzymatic methylation normally has a hydrogen deprotonation, but the 5-fluorouracil needs a fluorine to leave in the inhibitor mechanism, thus deactivating the enzyme.

RNA additionally modifies some of the bases within RNA, such as methylation on the exocyclic adenosine nitrogens and many modifications of uridine bases. It is hypothesized that these may promote transcription or alternative splicing, but not much is known.

Unlike DNA folding, RNA folding is much more versatile and has many local minimal structures. In order to combat this, we have the **polyelectrolyte effect** has shown that viscosity actually decreased when sodium chloride was added to a strand of charged polymers, since the polymer chain collapses. Similarly, double stranded regions of DNA and RNA are more stable at high salt concentrations. Thus, adding more salt could cause RNA to collapse to its global minimum structure.

**Example 28**

For the binding of protein and DNA, we have that $\Delta S$ (and thus affinity) is larger at low [salt] then at high [salt]. This is since DNA is negatively charged and attracts cations, while the positively charged protein is surrounded by anions. As the protein and DNA complexes, the anions and cations become more free, and thus $\Delta S$ increases. It is further found that $\log K_a$ and log [salt] give a linear plot, which allows us to quantify ion pairs.

Another structural feature of RNA is known as the **tetraloop**, where $\text{Base}_{i+3}$ makes a H-bond

pair with $\text{Base}_i$ while $\text{Base}_{i+3}$ and $\text{Base}_{i+2}$ are stacked together. This makes RNA somewhat more stable.

**Note 29**

With regard to RNA-complexation, we see that RNA-ligand protein complexation rates is significantly lower than protein-ligand and protein-RNA rates. (Ref: Herschlag: RNA **2017**, 23, 1745). We don't know why, however.

RNA is speculated to have played a major role in early life forms, as seen in many metabolic pathways and in catalysis. These observations led to the proposal of the **RNA World Hypothesis**.

> "One can contemplate an RNA world, containing only RNA molecules that serve to catalyse the synthesis of themselves."

Further evidence for this theory comes from an analysis of the ribosome structure (Ref: Steitz: *Science 289*, 920 (2000)). In ribosomes, the proteins involved were simply for stabilizing the negative charge of the RNA within ribosomes! The ribosome itself worked without the proteins. It itself is a ribozyme!

Current evidence shows that an RNA world existed about 3.8 billion years ago. There are plausible prebiotic routes to RNA, for example, from glycoaldehyde, glyceraldehyde, cyanamide, and cyanoacetylene (Refs: Sutherland: *Nature* **2009**, *459*, 239 and *ACS Chem. Biol.* **2010**, *5*, 655.) Sutherland has also proposed plausible routes to all the natural amino acids, from $H_2S$, HCN and UV light (Sutherland: *Nat. Chem.*, **2015**, *7*, 301.) Another unified synthesis of purines and pyrimidines is given in Carell: *Science*, **2019**, *366*, 76. Further, prebiotic mechanisms to copy RNA also exist (Ref: Szostak: *Angew. Chem., Int. Ed.* **2019**, *58*, 10812). We'll expand on this in next lecture.

# 6 Lecture 5: February 18, 2020

Today's lecture is again given by Prof. Raines. First of all, we have a take-home quiz, due Thursday, and standard MIT rules apply.

We'll start off by talking about plausible prebiotic routes of RNA replication. One of the first methods proposed was that an 2-methylimidazole moiety connected to a phosphate would result in an activated mononucleotide that could facilitate replication, but measurements showed that this was a slow reaction. However, if there are two phosphoryl groups connected to the imidizole (respectively on the 1 and 3 positions), then the system becomes more organized and more electrophilic, resulting in a much faster reaction. An even faster reaction happens when the methyl group is replaced by an amino group. Essentially, this imidazole moiety acts as an catalyst to promote replication from a template strand. Crystal structures of this process are given in the above Szostak reference.

We'll now take a step back and look at phylogenetics. Previously, we categorized species simply by how similar they looked. However, in 1965, Pauling proposed that DNA sequnces could be an innate clock that can be used to infer phylogeny. (Ref: *J. Theoret. Biol.* (1965) **8**, 357-366). Another relevant paper showed that most changes in nucleic acid sequence are "neutral (not changing the amino acid sequence)," and this allows for another method of monitoring this molecular clock. (Ref: *Nature*, **1968**, *217*, 624).

Original efforts in sequencing were based on small vertebrate proteins, such as hemoglobin and ribonuclease, which is common to all vertebrates. In the late 70's, Woese instead proposed to use the DNA that encodes an organism's rRNA, since ribosomes are present in all life and rRNA is homologous, and also changing very slowly, allowing more ancient comparisons. As a result of this classification, the phylogenetic tree was instead revised into three kingdoms of bacteria, archaea, and eucaryota. (Ref: Woese: *PNAS*, **1977**, *74*, 5088).

**Note 30**

In order to clear up some ambiguities, we'll define a few words in chemical biology. (Ref: Burgess: *Protein Expr. Purif.* **2016**, *120*, 106). Here are some word pairs/triplets that are somewhat misused:

- mutant - an altered organism or gene (NOT protein)
- variant - an altered protein

- homologous - evolutionarily related (yes or no answer, not % relation, etc)
- identical - the same (% relation is okay)
- similar - resembling (% relation is okay, with a definition of similarity)
- Ref: *Cell.* **1987**, *50*, 667

- *in vitro* - outside of a living organism
- *in cellulo* - within a cell taken from an animal
- *in vivo* - within a living organism

- enzymatic - by an enzyme (ex. enzymatic reaction)
- enzymic - of an enzyme (ex. enzymic residue)

- cytoplasm - everything in a cell except in ncuelus
- cytosol - the solution within the plasma membrane; the soluble portion

After this interlude, we'll now talk about **SELEX**, the Systemic Evolution of Ligands by Exponential Enrichment. It is a method for making **aptamers** (a short single-stranded nucleic acid that binds to a specific target) for target ligands, first discovered by Shoztack and Gold in 1990.

SELEX involves first making a library of ssDNA/ssRNA, adding the ligand on resin, washing away the unbound sequences, and amplifying the bound nucleic acids with PCR. This process is repeated many times, and what eventually emerges is a 'winner' that can bind the ligand the best.

Now, we've developed even more derivatives in the SELEX library, such as 2'-fluoro and 4'-thio nucleic acids, as well as other derivatives such as boranophosphates.

Another advancement is in using a live mouse in order to find binding to organs of interest. A random RNA pool is injected into a mouse, its liver is harvested, and the RNA is extracted and injected into another mouse. This is repeated similarly to SELEX and ultimately allows us to target diverse molecules.

One application of SELEX is in the of development of 'RNA GFP,' where aptamers were made to bind to specific RNA sequences. The main small molecules, HBI and DMHBI, were able to bind to the aptamer termed 13-2 RNA, selected from a library of $5 \cdot 10^{13}$ molecules. This resulted in green fluorescence. Further, other molecules have been found to bind to different aptamers, and each gives off a different characteristic color, anywhere from red to blue.

**Note 31**

What has been found is that DFHBI has been found to be the optimal ligand, with the optimal aptamer being 84 nucleotides and having binding at a G-quadruplex. The quantum yield $\Phi = 0.72$.

Though this is a chemical technique, in 2002, natural aptamers termed **riboswitches** (mostly based on RNA) have also been discovered. (Ref: *Nature* **2002**, *419*, 952). These riboswitches have been found in bacteria, archaea, fungi, and plants (but not animals!), and can bind to coenzymes, nucleobases, amino acids, ions, sugars, metabolites, etc.

**Example 32**

One interesting riboswitch is in the binding of fluoride ions. This is used by an extremophile when fluoride concentrations are too high, and at first, it doesn't make sense - how does negatively charged RNA bind negatively charged fluoride? The RNA actually incorporates three magnetsium ions that create order within the RNA, and these magnesiums are the ones that bind fluoride.

An interesting observation amongst DNA sequences is **Szybalski's rule** - the coding strand has more purines than pyrimidines. One possible explanation is that ancient organisms needs to have either many pyrimidines or purines in order to prevent double-stranded regions from forming in single-stranded DNA.

# 7 Lectures 6-8: February 19, 23, and 25, 2020

I was absent on these days, so no notes here.

# 8 Lecture 9: March 2, 2020

Today's lecture is by Prof. Imperali. She starts off by reminding us that the second PSET will be much harder than the first PSET, and will require a lot more time than simply one night of work. Additionally, lab experience will be very helpful in proposing methods/experiments for the PSET, especially with regard to the freshmen taking the course. Also, our second quiz will be passed out today.

We're going to continue our discussion of proteases, from the last few lectures. If mRNA makes a protease, then "all hell would let lose." As a result, proteases are expressed in an inactive form, primed for activation by an enzyme that modifies the protein with methods such as phosphorylation or carbohydroxalation. These are called **zymogens**. However, this doesn't tell us about the scope of activity of this protein. This is why we need **activity-based protein profiling (ABPP)**, which allows us to probe the protein at a functional level. These probes are typically tripartite, with a reactive group for labeling, a linker with substrate-type recognition, and a tag for visualization or a recovery from complex mixtures.

These 'reactive groups' are often developed specifically for a specific type of amino acid in the active site of the protein. For example, cysteine and series proteases react quite well with halomethylketones, where the histidine interacts with the halomethylketones. Another electrophile, suitable for serine proteases and esterases, uses a fluorophosphonate, in which the serine attacks the phosphorous and displaces the flouride. A third electrophile, an 2,3-epoxyester, is suitable for cysteine proteases, where the cysteine is modified.

**Note 33**

When using a halomethylketone, it is somewhat surprising that it is the histidine that reacts with the halomethylketone, rather than the cysteine, despite the fact that cysteine is a 'soft' nucleophile. Some explanations have been proposed to explain this.

Also, chlorine or bromine cannot be used in the fluorophosphonate labelling, because water would hydrolyze it much quicker than the protease!

**Example 34**

In the paper "Activity-based protein profiling for biochemical pathway discovery in cancer," ABPP was used to evaluate the activity of inhibitors against various non-agressive and agressive cells in various types of cancer. Based off a gel, we can tell that the inhibitor is able to irreversibly bind to cancerous enzymes (see paper for image).

**Remark 35.** *Some people suggest that a 2D gel should be ran to perform a similar experiment. Prof. Imperali says that anyone who has ran a 2D gel would never tell someone else to run one, since they are not easily reproduceable.*

**Example 36**

Another method for labelling is to use a **biotin affinity handle**. Biotin has an extremely high affinity for streptavidin, and elution through a column filled with streptavidin-beads allows for purification of the protein. Then, treatment with trypsin and a further mass spec analysis allows for identification of the protein. This even allows for separation of mixtures of proteins, since we can compare it with a protein database!

**Example 37**

We can also use ABPP to not only identify irreversible inhibitors, but also reversible inhibitors! This uses a differential analysis, by first binding the reversible inhibitors, then the ABPP reagent. The ABPP reagent binds to the proteins that don't have reversible inhibitors attached, and a gel (versus without inhibitors) reveals which inhibitors were successful. The advantage of this would be to check for cellular avaliability and labelling! (There are many inhibitors with high affinity that actually have no biological activity, since the inhibitor never gets into the cell).

**Question 38.** *Our ABPP examplese mostly relied on serine and cysteine residues, which are decently reactive and can react with our electrophilic reagents. However, how do we label proteins which have few nucleophiles at the active sites?*

For example, we have aspartyl proteases, which has two carboxylic acids at their active site. These carboxylic acids aren't very nucleophilic, and there are too many interactions between the side chains, water, and the actual substrate. If we attempt to label the nucleophile, we would end up labelling water! The solution is to change the linker in the ABPP to instead be a mimetic that interacts with the protein, since we don't have the ability to label the active site. These mimetics are commonly dipeptides, and the inhibitor has a extremely similar structure to the actual substrate. However, the amide of the scissile bond is instead replaced with a methinine hydroxyl, and are extremely potent.

**Example 39**

The HIV protease is an aspartyl protease, which is a homodimer in which the Asp25 and Asp25' residues form the active site of the bond. The HIV protease is able to cut up the polyprotein of HIV into many active proteins, which allow for the proliferation of HIV. Inhibitors use a mimetic extremely similar to the dipeptide mimetic, and many companies created drugs taking advantage of this. "It wasn't that they were brilliant; they were just copying nature."

For in-cell studies, we often just use the dipeptide mimic, that is attached to an alkyne, since sometimes larger molecules are hard to get into cells. Then, after lyzing the cells, we use click chemistry (the CuAAC) to attach a biotin-azide that allows for subsequent characterization. These are bioorthogonal, since there are nearly no alkynes in natural products and virtually no azides.

Another method for protein degredation is **ubiquitination**, in which ubiquitin attaches to the protein to be degraded, tagging it for degredation with the 7-fold symmetric **26S proteasome**. First, ATP is used to phosphorylate the ubiquitin, and the ubiquitin E1 enzyme, which contains a free sulfhydryl group, displaces the O-AMP. Then, the E2 enzyme causes thioester exchange, and E3 ligates the ubiquitin to the protein target. More and more ubiquitin is added to create a polyubiquitin product with the same procedure, and these polyubiquitin chains mark the protein for destruction. The proteasome uses 4 interdependent parts - a capturing mechanism, a deubiquitination system, an ATP-dependent unfolding protein, and a proteolyze. This proteasome doesn't care for the identity of any of these proteins; it just cuts them all!

This proteasome allows for some new drug discovery, specifically the **Proteolysis targeting chimeras (PROTACs)**. We start with the E3 ligase, which is bound to E2 with the ubiquitin that is ready to transfer the ubiquitin to a protein. We use a drug that has a ligand that binds E3, as well as the target protein, with a linker. This promotes ubiquitination of the target protein, which signals it for destruction. We don't even need a whole equivalent of the drug, since it can just get reused! However, if we have too much of the drug, then we have many open-ended linkages, and activity is actually reduced. However, the problem with these drugs is that it needs two binding units and a linker, and it is hard to get it into actual cells.

**Remark 40.** *In nature, there is a similar process, but with less specificity. Ubiquitin commonly binds to misfolded proteins which have different structures than the correctly folded protein, most likely through hydrophobic interactions.*

# 9 Lecture 10: March 4, 2020

Today's lecture is again given by Prof. Imperali. She starts off by mentioning that LigPlot (required for the PSET) has some issues on macs; she says to just email her if you cannot get it to work. Today's lecture will be quite a few topics smattered together.

We'll continue our discussion of PROTACs from last lecture with a small quote:

> "I like Joe Biden but I didn't vote for him. I like PROTACs but I don't vote for them either."

PROTACs have a molar mass of about 700-1000 Da, which makes them hard to get into the body, mostly through intravenous, intra-peritoneal, or subcutaneous routes. However, once they are delivered, then they behave like small molecules, and have intracellular targets. They have been found to show good activity; for example, 90% degredation of Brd4 has been seen in tumors.

Now, we're going to move onto N-terminomic analysis, which as the name implies, are methods for evaluating the activity of proteases through analysis of non-native protein N-termini. One of the methods of analysis is called **Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC)**, which isn't specifically for N-terminoic analysis, but rather just for studying cell pertubations. We use different lysine residues, light and heavy lysines ($^{12}C/^{14}N$ vs $^{13}C/^{15}N$), such that each heavy lysine residue is 8 Da more than the light variant. We grow cells in both light and heavy medium, and this environment will propogate to all the proteins in the cells. We also perform our pertubation on the heavy variant. After lysing the cells, we combine the heavy and light experiments, treat it with trypsin, and use LC-MS to look for peaks that are 8 Da apart. The ratios of the peaks then tell us about the effect of the pertubation. Then, bioinformatic methods can be used to put back the protein sequence.

**Note 41**

Deuterium isn't used in these isotopic labelling experiments since they different biochemical properties, due to the kinetic isotope effect. On the other hand, the other amino acids can also be used, but Lysine is preferred since it can be easily grown and biosynthesized from heavy ammonium acetate.

**Note 42**

Mass Spectroscopy is known not to be quantitative. However, when these two extremely similar proteins are sent through the MS simultaneously, then this problem disappears and the ratio of the peaks is actually quantitative.

Now, we're going to see exactly how to use SILAC to do N-terminomic analysis, called **Terminal amine isotopic labeling of substrates (TAILS)**. We likewise grow two groups of cells, one as normal and one with a pertubation. After lysing and fractionating these groups of cells, and then treating with trpysin, we perform reductive amination with formaldehyde (either normal formaldehyde or $D_2{}^{13}CO$) and $NaBH_3CN$, resulting in a N,N-dimethyl derivative where only the N-terminus reacts. This treatment is only done after the cell has grown, so it doesn't interfere with growth. The molecular weight increases by 6 in the heavy variant per every N-terminus. Then, we can use a mass spec to check for the ratios between M and M+6 peaks to tell the effect of our pertubation.

**Remark 43.** *The paper from which this technique originated is long and dense and horrible to read (Kleifeld, 2010, Nat. Biotechnol., 28, 281–288)*

**Remark 44.** *Before a compound gets a blessing from the FDA, it is not a drug, but rather just a probe.*

We're going to shift gears now and talk about **chemical genetics**, where we use small molecules to change protein function directly in live cells. We need these modifications to act on a scale of minutes, not too long, since otherwise it is impossible to attribute the effects directly to the small molecule. This differs from biological alternatives, such as temperature-sensitive mutations or gene modifications, which are often much slower but have high specificity. Chemical genetic techniques will have a range of specificities, depending on how good the probe is. They don't replace, but rather complement the biological techniques through being much faster. This time domain can be made even shorter with light-activated changes (known as optogenetics).

**Note 45**

In enzyme kinetics, we have the standard Michaelis-Menten competitive inhibitor kinetics model, which has parameters $K_m, K_I$, and $k_{cat}$. There is another method of analysis $IC_{50}$, which is the necessary concentration to inhibit 50% of protein function. This can be helpful, but is concentration-dependent and hence not useful for outside comparisons.

In biological genetics, we have both forward and reverse genetics. Forward genetics involves introduction of a random genetic mutation (through, for example, radiation, since we only want a few mutations) to a species, looking at the progeny, and choosing mutants that have distinct phenotypes. Then, we identify the affected gene. On the other hand, in reverse genetics, we introduce a selective gene mutation, develop the organism with the mutation, and we then figure out the role of the gene via comparison with the wild type organism.

Chemical genetics also has forward and reverse methods. In forward chemical genetics, we introduce a library of diverse compounds, and use a visible phenotype to separate out the an active compound. This compound binds to the protein, which can be separated through **affinity chromatography**, and hence the protein is responsible for the phenotype. In reverse chemical genetics, we instead use libraries of chemical compounds against a purified protein. After identifying this ligand, we add it to developing organisms, and use phenotypic changes to determine the protein's function. This allows us to find the function of a certain protein. This process, in fact, can be automated!

If we have compounds, then we can apply the methods of chemical genetics. But where do we find these compounds?

We have a few methods:

1. Rational Design - where we make analogs of native ligands, if they are known and have an understood mechanism
2. In silico modeling - where we use computers to predict binding with "hot spots" on the protein, followed with synthesis and measurement.
3. Compound libraries - where we make small molecules via synthesis, or from natural products and known pharmaceuticals. However, there are oftentimes undesirable compounds, which have 'bad' properties like reactivity, unstablility, flourescence, insolublility, toxicity, etc. which have to be removed from the library. Another property we want to avoid in compound libraries is **PAINS (pan assay interference compounds)**, which is when molecules aggregate and give many false positives in assays. These can be removed by adding 0.1% detergent.
4. Fragment-based screening - where small fragments of known inhibitors/drugs are screened, and is followed by chemical optimization.

**Note 46**

Antibiotic Discovery via deep learning has been a rapidly developing field (see Stokes, 2020, *Cell*, 180(4), 688). There is also another paper on using AI to find coronavirus inhibitors. However, these aren't too successful.

For the compounds that we end up deciding on, we have a few desirable properties, such as the easy of synthesis, cell permeability, oral avaliability, non-toxicity, and non-degredation.

**Remark 47.** *There has also been a* **rule-of-five** *analysis by Lipinski, which puts out properties of 'good' drugs, that ultimately ends up being useless.*

# 10 Lecture 11: March 9, 2020

Today's lecture is given by Prof. Raines, about protein folding. One of the first methods developed to identify folding was based on H/D exchange. First, a folded protein is denatured in Guandium·HCl, in $D_2O$. This replaces all the amide protons with deuterium. Then, when diluted, the protein refolds. Some regions of the protein fold fast, and the deuterium is incorporated into intramolecular hydrogen bonds, while the slow-folding ones get exchanged with protons before folding. These fast-folding regions can then be identified with either NMR or mass spec.

On the other hand, if we want to perturb folding, we can use some isosteres, defined as follows:

**Definition 48**

An **isostere** is a molecule with the same number of atoms and valence electrons arranged in the same manner; for example, amides, thioamides, and esters.

Each isostere has different properties - for example, a thioamide has higher cis-trans rigidity and better H-bond donor, but weaker H-bond acceptor. These differential interactions can ultimately change protein folding patterns, and these derivatives can be installed and tested by **protein ligation** (covered later in the course).

In general, we have the **Hofmeister series** that gives the effect of small ions on protein denaturation. The series mostly follow the hard-soft scale, with hard ions being **kosmotropes** and soft ions named **chaotropes**. The hard ions increase protein stability, while the soft ions decrease protein stability. Since soft ions do not have strong interactions with water, they start to become attached to peptide bonds, which disrupts the folding framework. On the other hand, hard ions are attracted to water more than to peptide bonds, and prevent water from interacting with the peptide bonds, increasing their stability.

Therefore, to denature protiens, we want to use a combination of two soft ions - 4M guanidinium thiocyanate. We also want to add a bit of $\beta$-mercaptoethanol (0.1 M) to break disulfide bonds. This combination allows us to degrade ribonucleases and isolate RNAs.

**Example 49**

Let us identify the factors increasing protein stability in RNase T1. By making specific substitutions, we can test the strength of coulombic interactions. Making a salt bridge (for example, by introducing a lysine next to a glutamate) contributes about 1 kcal/mol of stability. Similarly, removing a salt bridge loses about 1 kcal/mol of stability. By doing these substitutions, there are a few sources of instability - namely, loss of enthalpy from hydrogen bonding, loss of configurational entropy, but also increases in stability - increase of enthalpy from coulumbic interactions, and increase of entropy in liberated water. These interactions approximately cancel out, so that the magnitude of each interaction is not large.

We can also see the effect of charge on proteins - in general, we want the charge to be as close to neutral as possible. Each protein has its highest stability at the **isoelectronic point**, where the net charge is zero. In RNase, this has an effect of about 3 kcal/mol. This is since the charges essentially balance out and have paired coulumbic interactions.

**Remark 50.** *The isoelectronic point is the most likely pH for a protein to crystallize out.*

We can also add ligand mimics (transition state analogues) that are similar to the protein's active site, since active sites generally work by stabilizing these. For example, adding 0.2 M $Na_2HPO_4$ will increase stability of RNase T1 by about 4 kcal/mol, since RNase targets RNA molecules with phosphates!

Disulfide bridges also are quite prevalent and has a major impact on the stability of the protein. However, most of this is due to entropic factors - essentially, an entropy cost is already paid with the disulfide linkage. The disulfide bonds already decrease the folding conformational entropy prefolding, such that the magnitude of change in conformational entropy after complete folding becomes smaller. This increases stability by about 3-6 kcal/mol, depending on if the disulfide bond is over a short or long string of amino acids.

Finally, temperature has a profound effect - as temperature increases, the protein becomes less stable. Further, the stabilizing hydrophobic effect (an entropic effect) also decreases in magnitude. This follows directly from the Gibbs Equation.

Now that we've talked quite a bit about factors that contribute to protein folding stability, the question becomes - what are the effects of protein misfolding? One of the key discoveries are that proteins can become infectious diseases by themselves - a fact shown over and over again. This is discussed further in Soto: *Trends Biochem. Sci.* **2011**, *36*, 151.

The main hypothesis for these proteins is the **amyloid hypothesis**, which states that misfolded proteins form oligomeric amyloid fibrils which ultimately become insoluble and creates plaques and aggregates. The main method for this aggregation is through the beta sheets, which instead of folding upon each other rather just simply stack: **Griffin**: *J. Am. Chem. Soc.* **2016**, *138*, 9663. In fact, it is hypothesized that the fibril aggregate is the most stable form of the protein, which is kind of scary!

How to we counter this aggregation? One method is to use peptide mimics, for example, methyl peptides (**meptides**) that cannot hydrogen bond with another layer. Though it removes the problem of making the insoluble aggregate, it promotes accumulation of a neurotoxic soluble aggregate, which is still bad. In fact, it is even hypothesized that the solid precipitate is not itself neurotoxic, but rather just a reservoir for the neutrotoxic soluble aggregates.

Another method to deter aggregation is to use a stabilizer of normal protein folding. The main protein that has been used for development is **transthyretin**, a human plasma protein that transports thyroid hormone (thyroxine)as a 55-kDa homotetramer. The mechanism for aggregation is first unfolding the tetramer, then aggregating the monomer, creating neurotoxicity. This is thermodynamically favored due to precipitation.

However, some substitutions on the protein structure can differentially affect the aggregation rate. These single amino acid substitutions can drastically reduce (or increase!) the aggregation rate - the most impactful is Thr119 $\rightarrow$ Met119.

Now, how do we do this without protein modifications? Well, the protein tetramer binds to two molecules of thyroxin, and this contributes to the tetramers' stability through ligand binding (see above). As a result, competitive inhibitors such as **tafamidis** have been developed to take advantage of this property to prevent aggregation. The body does not need large amounts of thyroxine transport, so this method is actually pretty effective!

**Note 51**

This drug was approved in 2019, the first one that stops protein aggregation!

A third strategy for destroying aggregates was based on a racemic mixture of BRD4780 that remove aggregates of the protein MUC1 (Ref: Greka, *Cell* **2019**, 178). The mechanism is not clear, but it seems to send the aggregates to a lysosome, and this new method has led to significant discussion.

# 11 Lecture 12: March 11, 2020

Today's lecture is quite nonstandard, due to the recent coronavirus outbreak. The main updates - lecture videos will be posted on the Lecture of the actual lecture; there is now a forum on stellar; PSET 2 is due whenever (sometime during Spring Break); the other PSETs will be shortened; grad student presentations will still happen and be recorded; the final project is yet to be decided, but may be a review, wikipedia article, or NSF grant proposal. Also, the class was roughly split into two and assigned to the TAs.

Shifting back to chemical biology, we will talk about **optogenetics**, or the use of light to cause transformations within cells. We generally want to use light with more than 360 nm because otherwise DNA will be damaged. Other factors to consider are the efficiency of conversion, reversibility, and potential side product toxicity. The depth of penetration depends on the wavelength - generally, as the wavelength increases, the amount of tissue penetration increases as well. Red light has a penetration of about 5 mm, while violet light has a penetration about 1 mm. If cells are made in culture, then essentially any wavelength will be able to cause changes in cells. Additionally, we need the entity to be able to be delivered into or generated inside a cell, and have the kinetic parameters matching the observed phenomena.

There are two major classes of light-activated molecules - cages and switches. Cages are biological molecules that have a labile group that can be removed with light, usually irreversibly. Switches instead switch conformation when irradiated with light, such as azobenzenes that switch to the cis form. These azobenzenes will make the subsittuents much closer, but eventually switch back to the trans form. As a result, these tranformations are reversible.

**Remark 52.** *Organic chemists don't really like the 'cage' nomenclature, in contrast to biologists. They would prefer 'photolabile protecting group.'*

Regarding cages, these molecules have a common structure - with an otrho-nitro benzene ring connected to a biologically active leaving group. There are many common caging groups, including CNB, NPE, DMNPE, DMNB, CMNB, and NP, each with their own advantages in uncaging rate, quantum yield, inertness of by-product, solubility, and wavelength absorption. The mechanism is very similar for all, where the leaving group becomes converted to a carbonyl and the nitro group becomes converted to a nitroso group through a Norrish Type II reaction. There are many ways to get these caging groups into, including surface residue modification, reactive cysteine modification, or thiophosphoroylation.

**Example 53**

A azobenzene has been used to regulate $K^+$ channels in neurotransmitters: see Shaker Channel Nature Neuroscience 2004, 7, 1381-1386, Trauner.

There are many classes of switches. One is a simple cis-trans isomerism, found in light-sensitive ion channels. (Historical Commentary: Dieseroff NATURE NEUROSCIENCE 2015, 18) There can also be LOV domains (Light-oxygen-voltage-sensing domain), coming from cysteine binding to flavin upon light irradiation.

**Remark 54.** *All these optogenetic techniques look great on paper; however, these are often extremely unreliable.*

**Note 55**

This part of the lecture was supposed to take 15 minutes, but it took an hour and 15 minutes. Thus, most of the second part of the lecture, on fluorescence, will be recorded.

There are MANY fluorescent techniques. There are some useful resources - mainly, Raines, ACS Chem Biol. 2008 and 2014, and "Principles of Fluorescence Spectroscopy." This book is "very very valuble" and human readable and a very good resource. We'll continue with this next time.

# 12 Lecture 13: March 30, 2020

The last week of classes was cancelled, and this is the first online lecture, post-spring break. Any questions can be posted in the online class forum after each lecture, and will be answered within 24 hours.

We'll continue to talk about fluorescent methods, which have really grown to be one of the most prominent methods in chemical biology due to its wide applicability.

The general way fluorescence works is that a molecule in the ground state absorbs a photon and moves to a higher-energy excited state, and this excess energy is released as the molecule transitions back to the ground state. The energy released from fluorescence is lower than the energy put in, and hence the fluroescence spectrum has a higher wavelength peak as compared to the absorbance spectrum. This information is represented in a **Jablonski diagram**, with the $S_0 \longrightarrow S_1{}'$ transition being that of excitation, the $S_1{}' \longrightarrow S_1$ being relaxation/internal conversion, and the $S_1 \longrightarrow S_0$ being emission.

These fluorescent molecules have many properties, which can be tuned to suit one's needs:

- $\lambda_{maxEx}$ and $\lambda_{maxEm}$
- Quantum yield $\Phi$ (generally between 0.01 to 0.99). The brightness is equal to $\Phi \cdot \varepsilon$, the extinction coefficient.
- Excited state lifetime (usually ps - ns range, some metal ions in $\mu$s-ms range)
- Stokes shift - difference between $\lambda_{maxEx}$ and $\lambda_{maxEm}$
- Photobleaching - irreversible damage due to conversion to the triplet, that results in an irreversible change in the molecule that prevents it from being excited again. However, in some cases, for example in diffusion, photobleaching a small area will allow for tracking.

How do we actually put these fluorophores into action? We can attach them to some protein handles, such as malemides or halomethylketones, which we remember from earlier can target specific amino acids such as lysine or cysteine. These dyes can be modified, for example, by adding a double bond to change the fluorescence wavelength, or by adding sulfonate groups decrease hydrophobicity.

Now that we have our fluorescent molecule, we can characterize it by using a spectrofluorimeter, which uses an excitation source, a wavelength filter or monochrometer, and sends light of a specific wavelength through the sample. Then, with an emission filter and an emission detector, we can detect the wavelength of light emitted. This process can be simplified with a fluorescence plate reader (which can read 96, 384, or even 1536 wells) or fluorescence scanners (that can read gels or

blots). Another thing to watch out for is in fluorescence microscopes - due to scattered light, the images may not be clear. This is overcome with a confocal microscope, which uses a field $< 1\mu$m thick to do so.

One of the most prominent applications is in fluorescence activated cell sorting (FACS), which can separate a mixture of labelled cells based on flfuorescence. It uses vibrations to produce droplets containing individual cells, which are scanned with the fluorimeter, and then given a charge with current. These cells then pass through charged plates, where they can be separated by charge.

**Example 56**

One example of FACS is in the HTS screening of enzyme libraries. It starts with a library of genes, that produces different variants of enzymes. Addition of the genes to a water-in-oil emulsion results in the separation of the genes into discrete droplets, whose DNA/mRNA is translated into proteins. These enzymes can then bind to fluorescent molecules, and the FACS sorter will choose those with the highest intensity, and then the genes in those droplets are sequenced (C&B 2005, 12, 1291-1300).

In general, there are many applications of fluorescence. One application is in biomolecule labeling, that allows binding to a protein (either through a covalent or non-covalent molecule). Another application is to have a fluorogenic enzyme substrate, that can be used for assays. They can also be used as sensors for small molecules, such as protons are metal ions, if those cause changes in the flourescent wavelength. We can also use them directly to mark cells, in flourescent proteins.

A further application is in protease-substrate screening, of the P2 and P3 amino acids. We add a specific functional group to glass arrays, which our protein is then linked with via a chemical linker and a dot printer. Then, the sites where the protease cuts with highest efficiency (at P1 = Lys) will result in the most flourescence, which we can screen.

**Note 57**

Cysteine is often not used in the screening, since its redox properties can cause problems. This means that screening two residues will require assaying $19^2 = 361$ distinct proteins.

A further system is with a donor-acceptor fluorophore, known as **Flourescence Resonance Energy Transfer (FRET)**. Instead of emission of light directly from the donor, the energy instead emits from the donor fluorophore to the acceptor. For this to work, we need the emission of the donor to overlap with the excitation of the acceptor. Since the efficiency of the transfer is distance dependent, this allows us to measure distance on the cellular scale (10 Å to 150 Å)!!

**Definition 58**

We define the **Förster Radius** $R_0$ for a pair of fluorophores as the distance for which the energy transfer is 50% efficient.

**Example 59**

One example of the usage of FRET is in detecting the concentration of a zinc ion with a zinc finger protein. With a high zinc concentration, then two fluorophores (fluorescesin and rhodamine) in the protein become much closer, which results in a high acceptor:donor emission ratio. On the other hand, with no zinc, the emission of the donor is much lower.

**Example 60**

Another example of the usage of FRET is in **molecular beacons**, where there is a 18-30 base pair region which is complementary to the target sequence but not self-complementary, and a stem of about 5-7 base pairs that is. The ends of the stem contain a dye and a quencher (something that absorbs light, but does not emit light). When this molecular beacon binds to the specific DNA sequence, then the dye and quencher become far apart and fluorescence can be seen. This can be used in many applications, for example, in detection of mycobacterium tuberculosis resistance to antibiotics.

Now we're going to talk about flourescent proteins - something which essentially shaped chemical biology. The first flourescent protein isolated was the **green fluorescent protein**, isolated from jellyfish - the structure only had a protein and not a cofactor to cause fluorescence! This protein was chemically modified from a dimer to a monomer, and can be expressed through DNA recombination in many organisms.

The structure of GFP contains a protein basket made with many beta sheets, that surrounds a unique fluorophore inside the basket. The fluorophore is itself naturally made from a Ser-Tyr-Gly sequence (residues 65-67), in where the N of glycine condenses with the Ser-Tyr amide bond and is then oxidized by air to form a 5-membered heterocycle. This doesn't happen in any protein, but rather only when water is excluded (from hydrogen bonding) and the protein is in a specific conformation.

There were a few problems with GFP. One problem was that the flourescent signal had two peaks, and this was fixed by replacing Ser65 is with Thr65, resulting in breaking a H-bond and a much cleaner signal. Another problem was that GFP folded pretty slowly - once again, mutagenesis fixed this property.

A third problem was the actual color - how to we change it from a green color to any color we want? The answer was to change the aromatic amino acid and its environment, for example, changing Tyr -> Trp or Phe, which results in additional colors. But the most important advancement from this was that the cyan and the yellow variants formed a FRET pair!! This allowed the use of GFP for distance studies of other proteins.

Now we're hungry for even more colors. Going back to the jellyfishes, we isolated red fluorescent protein, that we were able to modify to create a rainbow of colors. Similar to the GFP, the RFP also suffered problems - it was isolated as a tetramer, and also had a slow maturation time. When the tetramer was broken apart (in both interfaces), there was minimal flourescence, and 33 mutations were needed to restore it fully.

How did we modify it and find out which specific amino acids to change? Some techniques were already mentioned earlier - the site-directed mutagenesis discussed earlier, casette mutagensis (changing multiple amino acids at a time), or error-prone PCR. However, the primary technique here is known as **gene shuffling**, where we take a family of related DNA sequences, fragment them into a pool of random fragments, and reassemble them into recombinants to be evaluated. (Curr. Op. Biotech. 1997, 8, 724-733)

Anyways, the end result is that we get a wide variety of colors that we can combine with FRET to observe cellular processes. Probably the most interesting obvservation is with FUCCI - Fluorescent Ubiquitination Cell Cycle Indicators - that allows us to visualize the entire cell cycle!!!!!! It's – amazing. (Cell 2008, 132, 487-498)

# 13 Lecture 14: April 1, 2020

Today's lecture is given by Prof. Raines, and we're going to talk about receptor-ligand interactions. The binding of a receptor to a ligand can be described simply by an equilibrium model as follows:

$$\mathrm{L} + \mathrm{R} \underset{k_d}{\overset{k_a}{\rightleftharpoons}} \mathrm{L}\cdot\mathrm{R}$$

where we define $K_d = \frac{1}{K_a} = \frac{k_d}{k_a} = \frac{[\mathrm{L}][\mathrm{R}]}{[\mathrm{L}\cdot\mathrm{R}]}$. The definitions of these constants are as follows:

- $k_a$ is the **association rate constant**, with units $\mathrm{M}^{-1}\ \mathrm{s}^{-1}$. It appears in the velocity of the forward reaction $V_a = \frac{\partial[\mathrm{L}\cdot\mathrm{R}]}{\partial t} = k_a[\mathrm{L}][\mathrm{R}]$, which has units $\mathrm{M}\ \mathrm{s}^{-1}$
- $k_a$ is the **disassociation rate constant**, with units $\mathrm{s}^{-1}$. For simple small-molecule ligands, $k_a \approx 10^8$, approximately the limit of diffusion. It appears in the velocity of the reverse reaction $V_d = \frac{\partial[\mathrm{L}]}{\partial t} = k_d[\mathrm{L}\cdot\mathrm{R}]$, which has units $\mathrm{M}\ \mathrm{s}^{-1}$.
- $K_d$ is the **equilibrium disassociation constant**, with units M. The affinities range from approximately $10^{-11}$ to about $10^{-4}$, with the tightest binding (lowest $K_d$) being between antibody-antigen complexes, and the loosest binding being enzyme-substrate complexes.
- $K_a$ is the **equilibrium association constant**, with units $\mathrm{M}^{-1}$.

**Note 61**

In the literature, capital $K$ always refers to the equilibrium constants, while the lowercase $k$ always refers to rate constants - they are not interchangeable. They are also italicized, in order to emphasize that they are simply variables.

**Remark 62.** *Chemists sometimes use $K_a$ to talk about host-guest interactions or binding affinities, but what is usually used in the literature is $K_d$.*

**Note 63**

In earlier times, there were many plots, such as the Scatchard plot, Eddie-Hofstee plot, Double-Reciprocal plot, and the Hill plot, which all linearized the data from receptor-ligand binding to be used in analysis. However, now there are many modern data-fitting methods that provide greater accuracy, so these plots are not used very often anymore.

For analysis, we'll define the bound ligand concentration $B = [\mathrm{L}\cdot\mathrm{R}]$ and the free ligand concentration $F = [\mathrm{L}]$. Some methods for obtaining $B$ and $F$ as follows:

- Filtration - simply filter a solution of ligand and receptor, such that the receptor cannot pass through the filter, but the ligand can. If we have some way of measuring the ligand and

receptor concentration (for example, absorbance measurements), then this can be used to quantify binding. However, keep in mind that the half life of such a ligand-receptor complex is usually on the order of milliseconds, which may make this methodology inaccurate.

- Gel Filtration - a mixture of receptor and ligand is passed through a gel-filtration column, and the receptor/receptor-ligand complex is eluted first due to the much higher molar mass. The ligand instead travels slowly through the column, and measurements of concentration allow us to find $B$ and $F$. This can also be used to identify tight binding interactions, if we use a library of ligands, and see which ones come through at the highest concentration with the receptors.
- Equilibrium Gel Filtration - first, saturate the column with ligand, then perform gel filtration. Measure the concentration of the receptor-ligand complex as it passes through the column, and what should be observed is a straight line, with one peak first (due to the extra ligands that the receptor pulls through) and then one trough (depletion of ligand at the end). The area under the curve reports $B$.
- Equilibrium Dialysis - have a concentrated solution of ligand and receptor, and add it to solution of the solvent. The ligand will mostly diffuse out, and the resulting concentrations within the dialysis bag can be used to calculate the parameters.
- Fluorescence Polarization - a small ligand is labelled with a fluorophore and is excited by polarized light. If the fluorophore is not bound, then there will be rapid rotation and depolarization. However, if it is bound to a receptor, then it cannot rotate very quickly at all, and hence the emitted light is polarized, which gives us a way to measure the binding parameters.

These techniques only apply when there is a significant size difference between the ligand and the receptor, but what if there isn't? Well, we have some more techniques:

- Affinity Chromotography - immobilize a receptor to a column, and add the analyte. The analytes that bind to the receptors are separated from the ones that don't. This is a qualitative assay.
- Immunoprecipitation - again, a qualitative assay. Here, we add an antibody that binds to the receptor. Then, we add bacterial protein A or G to make an antibody-receptor complex that is insoluble, and centrifugation will separate it, along with the ligand that it is potentially bound to.
- Surface Plasmon Resonance - have a metal surface that has either the ligand or the receptor immobilized (commonly the ligand), and have the partner flow through the top. We shine a single wavelength fixed angle beam of light onto the metal surface, and record its reflection. If the ligand binds to the receptor, then the light instead is absorbed (not reflected), and our detector can detect this. This technique, in fact, allows us to determine $k_a$ and $k_d$!

- Ultracentrifugation - this separates molecules based on their hydrodynamic properties (including but not limited to size). An analytical ultracentrifuge allows continual monitoring, and in fact a similar centrifuge was used to detect the $^{14}N/^{15}N$ differences that ultimately proved the semiconservative replication of DNA! This difference in isotope weight was able to work, and the differences between ligands and receptor are often greater.
- Electrophoresis - look for shifts in the gel separation between a probe and different molecules added. By adding different amounts of ligands, then we can measure $K_d$ to a factor of about 5.
- Isothermal Titration Calorimetry - "The Gold Standard." Measure the amount of heat released when ligands are added, and the data can give us $K_d$, $n$, $\Delta H$, and $\Delta S$.

**Example 64**

Rapamycin is a natural product that has immunosuppressant and antiproliferative properties, and it can bind to both FKBP12 and mTOR. mTOR was the target of rapamycin, but only in the prescence of FKBP12, which was shown with a column. (Sabatini: PNAS 2017, 114, 11818). It is known as a **chemical inducer of dimerization (CID)**, which more generally binds to two different proteins. These CIDs can allow for actual control of cellular properties, by linking two different proteins. There is one thing to watch out for - if we add too much of the CID, then the proteins themselves become saturated and do not link well, known as the **Hook Effect**.

# 14 Lecture 15: April 6, 2020

Today's lecture, and all the lectures from now on, will be given as a narration of the slies instead of a live lecture, since MIT has made it even more difficult to access on-campus resources. We're going to talk about bioorthogonal labeling, which ultimately allows us to track where and how molecules move within cells.

There are a few criteria to be aware of when designing a bioorthogonal label - selectivity, reaction biocompatibility, biological inertness, chemical inertness, accessibility, and kinetics. Additionally, we select these based on context as well - on what specific target our probe is targeted at.

**Definition 65**

Based on these criteria, we define a **bioorthogonal reaction** as one that can occur in the context of biomacromolecules and even in living systems, without interering with native biochemical processes.

One example of a bioorthogonal reaction is one we have already seen before - the Cu(I) catalyzed azide alkyne cycloaddition. To evaluate it, we look at the criteria mentioned before. The selectively is great, it is compatible with synthesis, and the kinetics are great! However, it may not be biologically inert, especially due to the copper involved.

Thus, some new methods were developed to not use copper instead. One example is to use a cyclooctyne as the alkyne partner, where the strain of the ring system promotes a cycloaddition, known fittingly as the **Strain promoted azide alkyne cycloaddition**. These molecules can also be modified to add handles to allow for fluorophore or affinity labelling.

**Remark 66.** *One of the commonly used cyclooctynes is known as BARAC, which is an abbreviation that stands for some of the functional groups. It was, fittingly, developed during the time of Barack Obama's presidency.*

**Note 67**

A quick note on rates - amine-carbonyl condensation has $10^{-4}$ rate constant, clearly not suitable, but it provides a basis for comparison. The CuAAC has a rate constant of about $10 - 200$, while the SPAAC only has a rate constant of about $10^{-2}$. However, a new methodology of reacting strained cyclooctynes not with an azide, but rather a tetrazene in an inverse-electron demand Diels-Alder cycloaddition, achieves rates anywhere between 1 to $10^5$. (all units in $M^{-1}$ $s^{-1}$, since these are second order reactions).

**Example 68**

Alkynes and azides are small functional groups, and can be incorporated into proteins by modifying an amino acid and making the amico acyl tRNA synthetases (AARS) attach the modified amino acid to the tRNA, and then the protein. Oftentimes, methionine is modified, since it is a relatively rare amino acid that is usually present in only about one or two amino acids per proteins. It is intuitive, and true, that knocking out the original cellular machinery for biosynthesis of methionine and increasing the concentration of the unnatural methionine will increase its incorporation into the protein! The most successful methionine analogue is to replace the SMe group with an alkyne, and this allows us to perform click chemistry on it! Another method involves adding an azide to alanine, and once again a CuAAC or SPAAC allows for visualization. We can use CuAAC here if we are studying bacteria, since they are relatively tolerant to copper, and can also use this to track the growth of mixtures of cultures!

Another technique used is known as **native chemical ligation (NCL)**. Here, we have two peptides and we ligate them together by changing one carboxylic acid to a thioester, and having a cysteine residue on the other N-terminus. Then, the sulfur of the cysteine undergoes thioester exchange with the thioester, and then the free amine displaces the newly formed thioester to join the two peptides together. Only the cysteine at the N-terminus can react, since we need the amide to form to make it irrevesible. Due to the requirement of cysteine, this is a chemoselective ligation, and is carried out under native conditions.

There are two methods for performing NCL. One method is to create the thioester from SPPS (solid phase peptide synthesis that can create peptides about 30 units long), and then ligate it with a much longer peptide, created with standard molecular biological techniques. Another method is to make the thioester of unlimited length with an Intein Approach, and ligate it with a SPPS made peptide.

For the first technique, we can make peptide thioesters with standard solid phase techniques. The peptide comes out protected, so we simply couple the free acid with the thiol, and then subject it to global deprotection to create our target thioester. The other protein can be synthesized by standard methods (with a ribosome), and having the first metionine being cleaved by a specific enzyme.

**Example 69**

One example of these techniques is in cell migration. The process is a cycle of extension, adhesion, translocation, and de-adhesion, and the central player is **paxillin**, which can interact with many parts of the machinery. It, in effect, is a scaffolding protein, that has over 40 phosphorylation sites. The pTyr31 is our target. We add a caged phosphate to the tyrosine unit to our peptide chemically (about 30 amino acids), and then using NCL to connect it to the rest of the 500+ amino acids. Then, by shining light, the photolabile protecting group falls off and our tyrosine becomes 'phosphorylated,' and we can see its impacts on the system.

To confirm this, we first make the natural part of the paxillin, with the key cysteine residue (Cys37) attached to a GST purification tag. Then, cleavage with a protease results in the loss of the tag, which we can detect. Then, adding our thioester reveals a change in molecular weight, which we can detect with a gel. We can also use an antibody that binds to a specific phosphorylated site to confirm our synthesis - if the photolabile protecting group is there, then the antibody cannot bind to it.

Now, we use total **internal reflectance fluorescence microscopy (TIRF)** to see this. We transfect the cells with GFP-vincullin, a focal adhesion marker. Then, plating on fibronectin allows us to track movement, after injection of the cells with the modified paxillin. There is no movement. However, once light is shined, the photolabile protecting group falls off, and now we know that the phosphorylation at Tyr31 is what causes cell movement!

# 15 Lecture 16: April 8, 2020

This native chemical ligation, as seen above, is pretty powerful. Now, we'll talk about the other variant of NCL - where we express a protein with a C-terminal thioester, and ligate them with an SPPS-made N-terminal cysteine. However, our previous technique (esterification then removal protecting groups) is clearly not compatible with living conditions, so how do we do it?

First we'll have a refresher on RNA splicing - the DNA is transcribed into normal mRNA, whose introns are removed, and then the mRNA is translated. In protein translation, a similar process occurs - where we have the protein sequence as an N-extein, then an intein, then a C-extein. The intein itself is marked with certain amino acids, including a cysteine at the end of the N-extein and Asn-Cys at the start of the N-extein. After **protein splicing** the N-extein and C-extein are joined by a native amide linkage (through an N to S shift, transthioesterification, Asn cyclization, and S to N shift), and the intein is removed. This process is *autocatalytic* (no other enzyme required) and *in cis* (another term for intramolecular, antonym *in trans* = intermolecular).

**Note 70**

These protein splicing mechanisms are often found mostly in single celled organisms. This can also lead some bacteria to be harmful through a process known as **intein invasion**,

So how can we exploit this to get a C-terminal thioester? Let's instead replace the Asn with an Ala residue - then, the Asn cyclization is unable to proceed. Addition of an excess of the thiol then results in a thioester being formed at the C-terminus of the N-extein. In practice, the C-extein often has an affinity tag attached to a carbohydrate framework, which allows for easy separation of the thioester.

Now let's look at an application of this semisynthesis technique. For more, see: Muir Accounts Chem Res 2009.

**Example 71**

We'll investigave the potassium ion channel (of bacteria), which is made with four identical subunits. The amino acids numbered 1 to 125 make up the membrane spanning region that is responsible for potassium selectivity.

The selectivity filter (residues 75-79) make the four ion binding sites, where the carbonyl groups of the protein (two from each subunit, for a total of 8) bind each of the metal ions. To investigate how this filter was so specific, this was investigated with NCL techniques.

The membrane spanning region was synthesized, with the amino acids 1-73 being synthesized with the protein splicing technique (with the aforementioned Asn/Ala variant) and residues 74-125 synthesized with chemical synthesis techniques.

The key modification was to change the 78-79 amide instead to the ester. Previously, in the wild type variant, we saw potassium ions almost waiting to go through the channel, with decent electron density even before S1 (through x-ray diffraction studies). However, with the ester there, the electron density of S1 is already lower, and there is much less before S1. The change ultimately results in the loss of a hydrogen bonding network and the ion binding site cannot bind water as well as the wild type, so the potassium ion cannot replace it. This, based on thermodynamic arguments, ultimately screws up the ion passage.

(References: JACS 2002, 124, 9113 and JACS 2006, 128, 11591: 'beautifully detailed')

Now we'll focus on enzyme-catalyzed labeling. The first method is with the SNAP self-labeling protein, based on $O^6$-alkylguanine-DNA alkyltransferase (hAGT), a DNA repair protein that removes the alkyl substitutent from an O-alkylated guanine. We attach the hAGT to our protein of interest. Now, if we alkylate our guanine ourselves with a probe, then the hAGT will take our probe with it when fixing the guanine, and thus allow us to attach it to the target.

There is also the CLIP variant associated with this. This causes hAGT to bind O-alkylated cytosine. If we combine these, then we can cause double-tagging by adding our probes attached to guanine and cytosine. This can also be used for real-time differentiation! Overall, our variant proteins SNAP or CLIP to these tags, and this offers complementary functionality to GFP.

The second method of enzyme-catalyzed labeling is with sortases, which are transpeptidases from gram-positive bacteria. The Sortase A was a transmembrane protein that recognizes a C-terminal sequence LPXTG (X can be any amino acid), and can sort the surface proteins. It is 206 AA long, but removal of the transmembrane portion results in a soluble protein 147 AA long.

**Remark 72.** *The sortases were named for the 'sorting hat' in Harry Potter.*

The mechanism is that sulfur of the Cys184 attacks the amide bond between T and G, ultimately removing glycine from the peptide. Then, the sortase causes the peptide thioester to react with a N-terminal GG or GGG-peptide. If we replace with our peptide with our unnatural molecule simply with a diglycine/triglycine tail, then we can ligate it to the C-terminus of the target. Similarly, we can ligate to the N-terminus via making an unnatural molecule with an LPXTG tail.

There are many sortases found in nature that allow orthogonal labelling. The one from *Streptococcus pyogenes* instead recognizes an AA-tailed N-terminus. If we ligate both unnatural molecules, then we can even make a FRET pair!

**Example 73**

This example comes from the Imperiali lab itself, in collaboration with the Griffith group. With a hydrogel polymer, LPXTG was linked to the polymer, and then epidermal growth factor was attached via sortase A. To quantify the amount of attachment, a large excess of GGG peptide could be added, resulting in the release of the growth factor. Finally, we could crosslink the gels to make a 3D polymer framework, whereupon cells were grown on and were broken upon addition of sortase, which lent itself to further analysis. This takes just 5 minutes!

# 16 Lecture 17: April 13, 2020

This lecture will be the third part in the lectures about modifying the genetic code, and we'll be focusing on site-selective mutagenesis. Recall from earlier that the only way we could modify the amino acid sequence with native machinery was to incorporate amino acids extremely similar to proteogenic ones, by having the bacteria grow in that medium. This method is both restrictive and uncontrolled. Now, we'll see some state-of-the-art methods for causing site-directed mutagenesis.

Recall from the genetic code that we have three stop codons - UAG (the Amber codon), UAA (the Opal codon), and UGA (the Ochre codon). In principle, we only need one stop codon, so we can reprogram the other ones to introduce the amino acids that we want. Most of the work done has been on the Amber codon, but some groups have also been working on the other codons, or even 4-base codons.

As we know, the protein is synthesized from a ribosome, made of rRNA and protein, which occurs in a ratio of about 2:1 in prokaryotes and 1:1 in eukaryotes (by mass), and is made of two subunits of size 50S and 30S (they combine to make one unit of 70S). The ribosome overall has a diameter of about 20 nm. What we can target in a ribosome are the enzymes known as amino-acyl tRNA synthases, which react with amino acid triphosphates to activate it, then binds to the tRNA, which then transfers the amino acid in the ribosome. These enzymes have quite diverse structures, with some being monomers and others being tetramers.

Thus, to get an unnatural amino acid, we can insert a stop codon into the mutant sequence, and engineer our tRNA to have an anticodon complementary to the stop codon and add the amino acid to it. Our modified tRNA can be synthesized with an RNA oligoligase and a dinucleotide-amino acid (direct addition of the amino acid does not work due to synthase enzyme specificity). Thus, if we use molecular biology techniques to modify the DNA, add the modified tRNA, and induce in vitro transcription, we should be to create modified proteins.

**Note 74**

Yeast tRNA actually works better than E. coli. tRNA, since it is less susceptible to editing. We further cannot incorporate D-amino acids, $\beta$-amino acids, or $\alpha,\alpha$ disubstituted amino acids, since the natural machinery is not suited for these.

Amide-to-ester switches were discussed earlier, and can be affected by NCL. This change can allow us to probe the electronic stability induced by hydrogen bonds in proteins. We can also use these unnatural amino acid mutagenesis techniques to affect the same transformation.

**Example 75**

In the T4 lysozyme, Leu39, Ser44, Ile50 were changed to leucic acid, leucic acid, and isoleucic acid. Leu39 and Ile50 are at the ends of an alpha helix, and Ser44 is in the middle. The replacement resulted in $\Delta\Delta G$ of 0.8 kcal/mol for the two on the end, and about 1.6 kcal/mol for the Serine, showing the impact of hydrogen bonding.

Another method for probing protein electrostatics is in using a **solvatochromic** molecule, a molecule that changes color based on its electronic environment, and measuring the changes in quantum yield. It can then be attatched to tRNA and incorporated into the protein of interest.

**Example 76**

The solvatochromin ALADAN amino acid was incorporated into varius residues in GB1. What was found based on the spectra was that A24 was on the surface, W43 is partially buried, L7 and F30 are completely buried. These changes of fluorescence can also help map out the protein structure.

We can also study ion channels in vivo, with the xenopus oocyte egg cell, which is decently large. We'll focus on an acetylcholine receptor, which needs to bind to acetylcholine for the channel to open. One method is to add two electrodes and measure the current across the cell, which increases as the concentration of acetylcholine increases. Another method is to use something called a patch clamp, which allows us to do single-molecule measurements in real time. Essentially, we use a glass capillary to get exactly one channel, and then this system is sealed off and can be measured.

**Example 77**

Let's try to perform this technique, with unnatural amino acids in the transmembrane protein. We microinject the RNA and the tRNA, and wait for about a day. Afterwards, we get our transmembrane protein around the oocycte. (note that the oocyte doesn't have these native channels, but they are still expressed!)

One study using this technique was performed on the Nicotinic Acetylcholine receptor, which causes addiction to nicotine. It has two identical $\alpha$ subunits, as well as $\beta, \delta, \gamma$ subunits. One of the $\alpha$ subunits has an acetylcholine binding pocket; however, probes showed that there were essentially only aromatic amino acids around that binding patch! The hypothesis was that this was a result of cation-$\pi$ interaction, rather than electrostatic interaction. How do we study this?

We can try to mutate the protein sequence to create measureable differences in the binding affinity. We replace Trp149 of the alpha subunit with fluoro analogues, and the more fluorines added, the more acetylcholine is needed to open the channel. This makes sense - the more fluorines there are, the more positive the ring is, and hence the cation-pi interaction is much weaker with the incorporation of more fluorines. The significant changes in binding from this structure-function analysis ultimately showed that the cation-pi interaction is very important.

The previous semisynthesis approaches worked fine, but we can do better. The semisynthesis was not very efficiency, and recoveries only make about ng-$\mu$g of protein. Additionally, the 'read through' vs termination ratio can be poor.

Thus, we should try to develop a better technique. The alternative is to exploit AARS/tRNA pairs from heterologous organisms and get them to do our bidding. The system that was found to have the most orthogonality was with *M. janaschii* AARS/tRNA in an *E. coli host*. The use of different organisms thus reduces the amount of cross-talk.

The general mechanism for doing this transfer is in six steps - selecting the orthogonal pair, verifying that there is no proof reading/editing, verifying that the anticodon loop can be changed, making a library of AARS for positive selection, making a second generation library for negative selection, and to loop back to positive selection with an unnatural amino acid. Ref: Schultz, Science, 2001, 292, 498 and Davis, 2012, Rev. Mol. Cell. Bio.

**Example 78**

In a Trp AARS, we can modify the 5 amino acids making the binding site to make the binding to O-methylated tyrosine about 100 times stronger than normal tyrosine. However, since the tyrosine is buried deep in the binding site, then there weren't many modifications that could have been made.

The proteinogenic amino acid Pyrrolysine (known as the 21st amino acid, 3-letter code Pyl/1-letter code O) solves this problem. D. Soll at Yale discovered the tRNA/AARS from archea had some good properties - the anticodon already matched to the amber stop codon, and shows high AA substrate promiscuity (already used for incorporation of more than 100 NCAAs). There are also other differences compared to other tRNAs - it has a stem of 6 base pairs instead of 5, a single base between (D and anticodon stems) and (D and acceptor stems), and has a three-base small variable loop. Finally, the pyrrolysine usage is not in a deep cavity, but rather just a groove - this gives us room for modification! (Liu, Biochem, Biophys Acta, 2014, 1844, 1059 is a review that shows many, many, many, possible incorporated modifications)

To move this technique in vivo, G. Church at Harvard genetically recoded E. Coli to not use the amber stop codon in the genome, and removed the stop recognition factor. This allows us to recode the amber codon to incorporate whatever we have on our tRNA, ultimately allowing us to incorporate these NCAAs in both prokaryotes and eukaryotes.

# 17 Lecture 18: April 15, 2020

Today's lecture is given by Prof. Raines, about the proteome and its diversity. A standard organism has about 20,000 genes, which can get translated into about 100,000 mRNA transcripts, that ultimately diversify to about 1,000,000 proteins, making the proteome much much more diverse than the genome. The amino acids that are commonly modified post-transcriptionally are glutamate and asparate, and those less often modified are the hydrophobic amino acids. Walsh has written a whole book and review on the topic: "Protein Posttranslational Modifications: The Chemistry of Proteome Diversifications."

Before our discussion, let's first define what we exactly mean by posttranslational modification as follows:

**Definition 79**

A **posttranslational modification** (PTM) is a covalent modification of a peptide or protein following its biosynthesis (not encoded by the genomie), that can take many forms - main-chain/side-chain modifications, as well as enzyme-catalyzed/non-enzymatic reactions.

Why are they useful? Many cellular components function based on these modifications, which can cause changes in structural and physical properties, activity, proteostasis, localization, and other interactions.

Currently in vitro, mass spectrometry is used as the premier technique, and before the advent of mass spectrometry, many minute changes such as oxygenation were unable to be detected. Western and Eastern blotting (respectively antibody and lectin based) can also be used in vitro. In vivo, antibodies can also work, but we can also use small-molecule chemical probes to react with the PTM.

**Example 80**

The first PTM that was detected was in the formation of insulin. It involved the usage of proteases to change pro-insulin into actual insulin, keeping the disulfide bonds already formed between these pieces. Another example is with the serpins, whose top loop is cleaved with a protease. Then, the top loop is able to interact with a system of beta-sheets in another part of the molecule, thus resulting in a 5-strand rather than a 4-strand sheet.

Another (complicated) example is the modification of Nisin, which undergoes a surprising 12 PTMs! The main modifications are the loss of water from alanine or threonine residues, forming $\alpha, \beta$ unsaturated carbons that can be attacked by cysteine to form thioethers.

A much simpler example is the conversion of N-terminal glutamine to pyroglutamate, reminiscent of native chemical ligation. This does not need an enzyme (though it can be catalyzed by glutaminyl cyclase), and the pyroglutamate residue can be cleaved by peptidase.

Not only can the main chain be modified, but an abundance of sidechain modifications are possible as well - namely, hydroxylation, lysine modifications, glycosylation, and phosphorylation. We'll be talking about hydroxylation and lysine modifications today, and the others later.

Hydroxylation of proline has been found to be very important in collagen, which is known as the most common protein and has even been found in dinosaurs. It involves (Xaa-Yaa-Gly)n repeating chains, where the glycine is required for the protein to form stable helicies. Commonly, Xaa = Pro and Yaa = Hydroxyproline (about 28% and 38% respectively).

Hyp also does not form beta-strands (which can lead to protein aggregates), such that its incorporation in the collagen can prevent such formation. Hyp is formed from hydroxyproline with the enzyme prolyl 4-hydroxylase, oxygen, $\alpha$-ketoglutarate, and Fe(II). This is a hard reaction - the mechanism has been revealed to be quite complicated, with the binding of the ketoglutarate, binding of the oxygen to the Fe, which then cyclizes to make a Fe(IV) species with the ketoglutarate. Loss of $CO_2$ then results in a succinate moiety and the active Fe(IV) catalyst. Sometimes, Fe(IV) can just revert back to Fe(III) without abstraction of a hydrogen from proline, and it needs to be reduced back to the active Fe(II) by vitamin C - hence underconsumption of vitamin C can result in less hydroxyprolyation (scurvy).

In 1973 (Ref: Prockop: Biochem. Biophys. Res. Commun. 1973, 52, 115) it was found that the hydroxyproline residue increased the melting temperature of the collagen triple helix by about 30 degrees, showing that it contributes much to overall stability. Another study (PNAS 2000, 97, 4763) showed that removing hydroxyprolines from a worm species resulted in significant structural

shortening.

The chemical reason for the increased stability of the hydroxyproline is based on the endo-exo pucker - the hydroxyproline residue has an exo pucker, thus reducing its $\phi$ angle by about 22 degrees, thus allowing it to form a more stable triple helix. The pucker can also be by addition of other groups, such as addition of methyl groups or halides. These allow us to control the specific specificity of our pepptide.

Another example in the modification of collagen is where we use a lysyl oxidase, that uses a quinine cofactor in the extracellular matrix. It oxidizes a lysine residue into an aldehyde, which condenses with another lysine in another chain, and thus results in cross-linking.

A third example is $\gamma$-Glutamyl carboxylase, which adds a carboxyl group to a glutamate residue alpha to the acid group. This enzyme uses an epoxide derived from vitamin K to induce this deprotonation of the alpha position, and acts on prothrombin and clotting factors, causing blood coagulation.

We also have many possible modifications with ubiquitin. As discussed earlier, ubiuqitin is a protein marker, that among other things, mark proteins for degredation. However, the ubiquination code is extremely complex, and can take a variety of forms, from monoubiquination to polyubiquination at different locations to formation of polyubiquitin chains. It involves a condensation between a lysine sidechain and the C-terminus of another protein. Thus, the more lysines there are, the more possibilities there are for ubiquitination results.

Though the ubiquitin code has not been fully unraveled, one method we can use involves the Staudinger reaction. Here, an azide is used instead of a lysine, and is incorporated with the techniques mentioned earlier. To couple this azide to ubiqutin, we use a ubiquitin equipped with a sulfuryl diarylphosphine, which is able to reduce the azide and hence form a peptide bond.

Another method is to use a delta-thiolyated lysine, which is able to react with a ubiquitin sulfuryl ester. This can also be protected with formaldehyde (forming a 5-membered cycle). Then, S-N acyl transfer and desulfuration leaves behind the requisite peptide bond. This process can be repeated many times to make a polyubiquinate chain (through Lys48).

We also have many non-enzymatic pathways for PTMs. These occur, for example, in the modification of histones, resulting in them packing DNA less efficiently.

One non-enzymatic reaction is lysine acylation, where the lysine sidechains spontaneously attack activated esters. Situins (an enzyme) can reverse this transformation. Aside from acyl groups, biotin, glucoic acid, and formic acid can also be added. This is also possible with Staudinger

ligation, which was mentioned above. Another enzymatic reaction is lysine methylation, where a cysteine reacts with $XCH_2CH_2NR_3^+ Cl^-$ where $R = H, Me$. The sulfur reacts with the halide and essentially makes a lysine analogue.

We also have non-enzymatic irreversible PTMs. One example is the Maillard reaction, that ultimately degrade glucose to make dicarbonyls that can interact with proteins to cross link. These cross-links have been seen to be present in all species, and our chemical synthesis techniques allows us to study them in depth. Another example is in the modification in the biosynthesis of GFP, as discussed last week.

# 18 Lecture 19: April 22, 2020

Today's lecture is given by Prof. Raines, and today's lecture will be about the human kinome, another extension of the proteanome. Humans have about 550 kinases, and about 30% of all human proteins are phosphorylated, implying that most of these protein kinases will have many protein targets. In addition to the 550 kinases, there are about 150 phosphoprotein phosphates, enzymes that remove phosphrylation. Thus the balance between these two classes of enzymes determine the relative degree of phsophorylation. It is also found that about half of the phosphosites are constratined by evolution, making them most likely functional. These phosphosites can give us significant information about protein function as a whole. Overall, phosphorylations turn proteins 'on,' while dephosphorylation turns them off.

In the phosphorylation process, a serine residue is able to attack ATP (with a Mg or Mn cofactor) after deprotonation, creating an ADP molecule and a phosphorylated oxygen. The phosphoserine side chain has two new pKas: 1.54 and 6.31, changing the charge by $-2$ at human pH = 7.365. These negative charges can promote new interactions, for example, with arginine residues or partial positive charges in $\alpha$-helixes [this is seen in insulin, where there is a large structural change after phosphorylation].

Though we can't directly introduce phosphate groups, we still have some other methods to mimic such a residue. Asp has been used before to mimic this, but it really isn't a good substitute.

Davis showed that a cysteine can be modified to a dehydroalanine by a desulfinating agent, where another thiol (or specifically a phosphothiol) can be introduced via a Michael addition (but stereochemistry is lost). Hackenberger instead modifies the cysteine (with retention of stereochemistry) by introducing an electrophilic disulfide, from which introduction of phosphite causes formation of a phosphothiol. Both of these methods make phosphocysteine.

Experimental approaches to cataloging phosphorylation are numerous: $^{32}$P radioactivity, MS spectroscopy, native chemical ligation and caged phosphates, engineered kinase/ATP, and kinase inhibitors.

In Engineered kinase/ATP pairs, these kinases often have a large 'gatekeeper residue' (F, I, L, M, T) that binds to the amino group of ATP. This helps with binding these, and rejects compounds like N-benzyl ATP. As a result, replacement of this gatekeeper residue with something like glycine will result in a 'hole' that makes room for the 'bump' made by Benzyl-ATP.

Usually, kinases are **pleiotropic**: one kinase phosphorylates many substrates. If we use a radioactively-labelled benzylated ATP, then only the products of the kinase variant are labeled, which can allow us

to recognize which residues are phosphorlyated by a specific kinase. These experiments, however, have to often be done ex vivo, since Bz-ATP is hard to get into cells.

Similarly, small molecule bumped inhibitors can fit in and inhibit kinases with holes, but not the normal kinases. This allows for shutting down certain kinases, and gives the complement of what we had earlier.

Now we're going to move onto cell signal transduction. As we know, cell signals are transmitted through molecules, and in many cases, these signals are transmitted through phosphorylation events. In general, a signal transduction cascade works through kinase activation, then transcription factor activtion, then promotion of DNA transcription, and finally a cellular response.

An understanding of cellular signal transduction is quite important to understanding exactly what goes wrong in cells. Cancer is a prime example. Though there are many factors that influence cancer (see Hanahan, Weinberg: Cell 2011, 144, 646), many signals that indicate cancer are based on kinases. There are both nontargeted and targeted therapies - chemotherapy, radiation therapy, small-molecule drugs that target specific kinases, etc.

The small-molecule ATP mimetics can inhibit the kinases and occupy either a DFG-in or DFG-out conformation (DFG is a tripeptide activation loop). There are also covalent and allosteric inhibitors. However, these often lack specificity, and further need to compete with the high intracellular concentrations of ATP.

**Example 81**

Chronic Myelogenous Leukemia is due to a chromosomal translocation between chromosome 9 and 22, and creaes a Bcr-Abl fusion protein which has increased kinase activity that causes leukemia. It proceeds in three phases: chronic, accelerated, and blast phases, but only the chronic phase was able to be stopped before 2001.

In 2001, Gleevec was approved as a therapy for CML, and was much more effective than the previous treatment, by inhibiting Bcr-Abl with high selectivity, and has only has some affinity for three other kinases. For Abl, the DFG-in state is the one necessary for phosphorylation, while the DFG-out state is for the inactive conformation. Gleevec works by binding to the inactive DFG-Out Abl kinase, and also binds to the Src kinase. These crystal structures are indistinguishable, yet Abl has a 2500x higher binding affinity to Gleevec. There were also no apparent sequence correlations between Gleevec-sensitive and Gleevec-insensitive kinase.

Why? The current hypothesis is that there is a large energetic pentaly for Src and other kinases to adopt a DFG-Out conformation. In other words, Abl is more flexible and able to adopt the binding conformation, rather than Src.

There is still a possibility of development of resistance, with substitutions in the Bcr-Abl protein possible (either direct or remote substitutions). Now, second-generation Abl inhibitors have been developed to overcome this.

# 19 Lecture 20: April 27, 2020

Today's lecture continues the theme of proteasome modifications - in glycosylations. The go-to reference for literally everything related to glycobiology is "Essentials of Glycobiology," which is actually avaliable free online!

In previous classes, carbohydrates are often just presented as an energy source, or a backbone for nucleic acids, but they actually have much more functions within so many cellular processes. There are many variants of sugars, from glucose to xylose to fucose to N-acetylneuramic acid, for a total of 9 main sugars in vertebrates. On the other hand, bacteria have over 750 sugars with a variety of diversity.

These sugars can form far more configurations as compared to nucleic acids and peptides, as it can make branched glycans from any of the avaliable hydroxy groups. There is a nomenclature for naming sugars, with colors and shapes depicting every sugar.

Protein glycosylation has many purposes, among them being helping during folding (often being co-translated), increasing activity, hindering proteolysis, helping with localization, and increasing interactions with other cellular compoents. As we can see, there are many many possible functions, leading to the following quote:

> "Protein Glycosylation - All the Theories are Correct." -Ajit Varki

There are significant new challenges with glycobiology - they are not template driven, they are hard to chemically synthesize, they can't really be detected, and there is significant heterogeneity.

In the cell, there are O-GlyNAc glycoproteins only. In the extracellular domain, there are glycoproteins (small amount of glycans) and proteoglycans (small amount of protein), as well as glycopolymers. N-linked glycoproteins (attached to Asn) and O-linked glycoproteins (attached to Ser/Thr) are the most common glycoproteins, though there are also others which are anchored to the cell membrane. Since most glycoproteins are outside of the cell, they are able to be used for cell communication.

The biosynthetic pathway to create a N-glycoprotein is one that starts with highly nonpolar polyprenol phosphate, from which activated sugar donors react to make a polyprenol-linked glycan. This is transfered directly to the protein, which has to have a Asn-Xaa-Ser/Thr **sequon** to recognize it. In bacteria, this process is very similar, though the sugar/protein identity and location is slightly different.

In eukaryotes, the glycoprotein is synthesized at the interface on the membrane between the cy-

toplasm and the endoplasmic reticulum. It starts with a phosphoglycosyl transferase, from which glycosyltransferases add sugars. In the middle of this process, there is a **flippase** that brings the sugar to the ER-side of the membrane. Then, a large oligosaccharyl transferase transfers the glycan to the protein. Then, glycosyl transferases and hydrolases can further diversify the glycans in many, many ways (this diversity varies by organism).

Viruses have tiny genomes, but can still hijack the glycosyl machinery. Coronaviruses have four structural proteins (spike, envelope, membrane, and nucleocapsid). The spike protein is the one that attaches to the human cell (through the ACE2 and TMPRSS2 receptors), from which it exploits the entire human proteasome.

When using cryo-EM or crystollagraphy techniques are used, the glycans move around significantly, so molecular dynamics computations are necessary to locate them. The proteins are coated with glycans, which look similar to human sugars, from which immune recognition is significantly harder. Further, it is also hard for proteases to attach to the protein.

We can use metabolic labeling in glycoproteins as well, but this is much harder than the variant for amino acids. We need to get the actual molecules in the cell, and also get them to be picked up by the cell's machinery, from which we can use a bioorthogonal ligation to attach a fluorophore or biotin affinity probe.

The azide group can be incorporated into the alpha position of the N-acetyl group, and the hydroxyl groups all need to be acylated to be integrated into the cell. Afterwards, the GalNAc salvage pathway replaces the anomeric hydroxyl with a hydroxyphosphate, from which it is replaced with a nucleotide and integrated into the new glycoprotein on the cell's surface.

Zebrafish have been raised with the modified galactose and a cyclooctyne derivative. A comparison of a control and a modified zebrafish reveals that the modified zebrafish's embryo development is much more significant. (Science 2008, 320, 664).

We can also change this to have two-color metabolic labelling! Essentially, we attach one dye, use TCEP (a quencher for azides), then repeat the process to essentially map out where each structure emerges during the embryo development process!!

We can even make this cell-specific. We make the glycan surrounded by the ligand-labelled liposome, which will deliver to specific cells only. This can be exploited in many applications, for example, in drug delivery for treating cancer. (Xie, JACS, 2012) Folate receptors are often present in epithelial-derived tumors, from which our specially designed liposomes can target.

# 20 Lecture 21: April 29, 2020

Today's lecture continues our discussion about glycobiology. When we survey glycan-protein binding sites, it is found that tryptophan (by far), tyrosine and histidine are the main amino acids found next to the binding sites. This is despite the low natural frequency of these aromatic amino acids in normal proteins. The proposed reason for this is aromatic-sugar interaction between the carbohydrate C-H and the pi-systems of the aromatic amino acid. Essentially, these aromatic systems are able to interact with the axial C-H bonds of the sugar rings, to essentially act as 'pincers.' (Hudson JACS 2015).

Polar residues such as Asp and Asn are also decently prevalent (these two are less prominent than the aromatics, but much more prominent than the others). These polar residues help increase specificity for sugar binding, while the aromatic rings help bind general sugars. Glu and Gln are less prevalent than Asp and Asn; this is hypothesized to be due to the greater degrees of freedom offered by the longer carbon chain resulting in less specificity.

**Note 82**

Despite all these interactions, these interactions are still very weak and the affinity coefficients are low.

There is also the phenomenon of 'multivalency,' where we have the proteins and glycans bind multiple equivalents, sometimes even being able to make cross-linked lattices. These increase the affinities significantly.

**Definition 83**

We define a **Carbohydrate Recognition Domain (CRD)** and a **Glycan Binding Protein (GBP)** or **lectin**, respectively as when a glycan binds to a recognition domain in one way, vs when a glycan binds to a protein entirely (through multiple domains).

The CRD has a affinity constant $K_d$, while the GBP has an **Avidity constant** the apparent $K_d$ (though this comes from non-equilibrium measurements). We can measure the association and disassociation rates in order to ultimately determine the relevant physical constants. Two common techniques are surface plasmon resonance (SPR) and biolayer interferometry (BLI).

There are many possible multivalency scaffolds - with factors such as flexibility, valency, matching, and spacing being all important factors. These scaffolds need to be developed with just the right amount of each of these factors in order to have specificity for the glycans.

These lectins often have a triple helix structure, which can even aggregate into a supramolecular structure to make higher order valency. This multivalency allows cells to 'roll around' (in leucocytes and selectins), almost as a 'velcro effect.' Another effect is in clumping of bacteria agglutination when exposed to a lectin, and this could be developed into a biosensor.

However, there are a few drawbacks to the aforementioned techniques to study glycan-protein interactions. Carbohydrates are hard to get in large quantities, are labor intensive, and is based on carbohydrate presentation. A **glycan array** overcomes all these difficulties, and this is similar to the protease arrays discussed earlier. We have a glass-slide where many glycans are adhered, from which a fluorescently-tagged GBP is added, which can be measured. This process can also be reversed to have lectins adhered to the glass slides.

A key limitation is in glycan binder technology, due to the diversity of bacterial glycans. As a result, efforts have been focused on **directed evolution (DE)** to generate new binders, which have found many applications (clinical, biotechnology, basic research, therapeutics, etc).

DE mimics natural selection, where we use many rounds of mutagenesis, and then select the best genes, which are amplified for use in the next round. It can be performed in vivo or in vitro, as long as we have a method of choosing which genes/proteins move forward.

**Example 84**

Lamprey antibodies have evolved to a completely different shape and structure than mamallian antibodies, and they have been seen to have good binding affinity to sugars. As a result, they have undergone DE through a process called **yeast surface display**. The yeast has Aga1p disulfide linked to Aga2p native surface proteins, and the relevant lamprey protein (attached to a labeller, encoded in a plasmid) is attached to Aga2p. For every yeast, we have 1 VLR variant, and we add a fluorescently labelled glycan. These can then be selected to take the best proteins/genes, which further undergo DE.

# 21 Lecture 22: May 4, 2020

Today's lecture is given by Prof. Raines, and we'll be going over some of the more recent advances in chemical biology: targeted therapies, siderophore-antibiotic conjugates, and antibody-drug conjugates.

Normal nontargeted approaches (such as chemotherapy) results in nonspecific toxicity and narrow therapeutic window, where the therapeutic dose is close to the maximum tolerated dose, and can damage organs. With targeted therapy, we essentially give a large window between the therapeutic dose (targeting just one organ) and the toxicity dose.

**Example 85**

Albomycin is a natural product that has a **siderophore** (something that binds a metal) and an antibiotic, which is able to cross the membrane and inhibit seryltRNA synthetase. This is a natural example of a siderophore-antibiotic conjugate.

Siderophores have quite a bit of structural diversity, where the best siderophore enterobactin has an association constant of $10^{52}$, binding iron so well such that iron is not released until the molecule itself is destroyed. This leads to the discovery of a drug motif, where a siderophore, linker, and antibiotic can be connected. The linker can both be cleaved or inert.

However, these conjugates have not been FDA-approved even 75 years after the discovery of albomycin, mostly due to low efficacy and potential microbial resistance.

Antibody-drug conjugates (ADC) have also seen development, designed as a "magic bullet" against cancers. These have a similar structure to siderophore-antibiotic conjugates, with an antibody, linker, and drug. Some of these ADCs have actually been approved for therapeutic use, starting from 1997.

**Note 86**

These antibody-drug conjugates are actually antibody-toxin conjugates, where the toxins used are much much more potent than previously discussed drugs. Thus, high specificity is needed in order to not damage normal cells.

The common ways to add toxins to the antibodies are through the lysine and cysteine residues. In cysteine residues, some of the residues are reduced and used for linkage, because the others are important for structural integrity. Bioconjugation techniques (as discussed earlier in the course) work well, with maleimide-based linkages being the most common.

One problem with maleimide-based linkages is the reversibility of the reaction, which can result in the toxin being released into the body. Instead, a methylamino group alpha to the nitrogen can be added, which promotes the hydrolysis of the maleimide that prevents the reverse reaction.

The toxins can be attached to the antibody in different ways. If we have a 'inert' linkage, then a lysosome causes unspecific degredation, which can result in a 'brute-force' approach to release the drug. If the linker is a hydrazone (for example, in mylotarg), then acid-catalyzed cleavage can instead happen. Protease cleavage is also possible.

Antibody-only activation is also possible, where binding to the antigen results in the immune system attacking it, but these effects are typically secondary compared to that of the toxin.

**Example 87**

Adecetris is an example of a successful ADC. The antibody Brentuximab binds to the CD30-antigen, which is preferrentially expressed in cancer cells. The toxin is auristatin E and is a derivative of a natural product Dolastatin 10, which is a pentapeptide.

Auristatin E was originally proposed to be attached through the hydroxy residue, but the possible esters made were too cleavable and the possible ether was nonreversible. As a result, they instead decided to attach the antibody through the N-terminus, through a carbamate and a long peptide-like linkage that is cleaved with a protease.

The drug works by preventing tubulin polymerization by binding to the tubulin moieties, which ultimately leads to apoptosis. The toxin is much more toxic than standalone drugs such as Taxol.

# 22 Lecture 23: May 6, 2020

Note: This is the last lecture for the course.

This lecture will be about macromolecular delivery to cells. Overall, about 80% are small molecules, where about 45% act from the outside and 35% act from the inside. Proteins make up the remaining 20%, with almost all of their activity through the outside of the cell.

There are very few proteins that actually get into cells. There are a few ways to do this - electroporation (ex vivo), microinjection (ex vivo), liposomes (not very developed), cationic lipids (ex vivo, and can contaminate cell), cell-penetrating peptides (CPP), and viral vectors.

Cell-penetrating peptides are often cationic, complementing the glycans of the cell which are often quite negatively-charged. The Tat protein (from HIV) was the first discovery of these cell-penetrating peptides, which results from its sequence containing highly positively charged RKKRRQRRR. Tat-based proteins can deliver many things through membranes and even through the blood-brain barrier.

Though it is not hard to get these cell-penetrating peptides into cells, it is hard to get out of the vesicle once it is inside the cell. Efficient new peptides, such as cyclic oligoarginine peptides and beta-peptides, have also been developed.

The oligoarginine peptide has a flexible backbone and a flexible spacer. This can be independently modified to make a rigid backbone and a rigid spacer. A arginine-attached proline-like oligopeptide (Z8, has both a rigid spacer and a rigid backbone) has quite high activity, and was not degraded as quickly as Tat. The enantiomer had the same activity, implying that this is a biophysical phenomenon.

Non-cationic cell-penetrating peptides are possible. For example, amphipathic CPPs have essentially a hairpin-like structure that can insert into the membrane. For example, Sanguinamide A is a natural product, whose intramolecular hydrogen bonds increase hydrophobicity and go through the membrane. Gramicidin A, another natural antibiotic, also does this, where it makes a giant helix that increases hydrophobicity.

Another application is in **arginine grafting**, where site-specific mutagensis changes negative amino acids to arginines. Doing this to GFP allows it to be taken up into cells, but still result in them being trapped in endosomes.

How do we fix this? One thing to do is to look at what has been done with previous small molecule drugs. Esters have been used in prodrugs to mask a carboxylate group, which would be degraded

by an esterate to make the actual drug. If we use a diazo compound to mask the carboxylates, then the protein itself can also be uptaken by the cell (even without endocytosis) and have the ester groups being degraded by esterase! (JACS 2017, 139, 14396)

In cells, the concentrations of protein and RNA is much much higher than the ones that we get in test tubes. We have significant 'crowding,' and macromolecules make up about 1/4 of the volume of the cell. As a consequence, diffusion is about 100 fold slower, and protein interactions are about 10 fold tighter, making antagonization much harder in cellulo. Further, protein folding and aggregation is more rapid and stable in cellulo rather than ex cellulo, due to entropic effects and enzymes.

To mimic these conditions ex vivo, we instead add crowding agents such as bovine and human serum albumin. Adding polymers such as polyethylene glycol also works to increase crowding.

**Note 88**

Viscosity does not necessarily mean that the cell is more crowded. Macroviscogens (such as proteins and polymers) do not affect the diffusion of small molecules (they can essentially ignore the matrix of the macroviscogens)! This is seen in the study of triosephosphate isomerase, which is inhibited by small molecule viscogens, but not by large molecule viscogens. This is an important point to consider when designing experiments.

In these crowded situations, we still have significant interactions between proteins that can find each other, with work done by Gunter Blobel. He hypothesized and verified that proteins had some signals attached that would mark its destination, through a protein channel (observed only 16 years later), ultimately leading to his Nobel prize in 1999 (Cell, 99, 557, 1999). It is also hypothesized that E. Coli has similar interactions.