GEORGE CHURCH: OK, ready. OK. So welcome to the third phase of our omics discussion. Someone pointed out that I should point out [? this shirt ?] says omics not comics. So we've covered genomics and last time transcriptomics, and today we introduce a very important all-inclusive subject of proteomics.

We'll connect it to last week's through the vehicle of focusing on motifs that are involved in protein interactions with the two nucleic acid macromolecules. So we're going to be covering, just as we introduced RNA omics with RNA structure, we're going to spend this entire class talking about protein three dimensional structure, how you get at it experimentally and computationally and its implications for the binding of small molecules such as drugs. We will in short order get to the scary pumpkin-like molecule.

So the connection to last week was this diagram showing palindromicity in three cases and a direct repeat in the fourth case. And I offered that this might reflect the symmetry of the proteins-- of the three dimensional structure of the proteins and the three dimensional structure of the nucleic acid and these symmetry elements would align. Now in order to introduce these symmetry elements and the possibility of having codes that you can at least program, even if they may have been tinkered about during an evolution, the question is to what extent can we get our hands on these kind of protein and nucleic acid motifs that interact.

In order to get at this issue of where there is a code-- and I just take this as one of the ways of dealing with the incredible complexity of proteins is to give this a theme that connects it to the last class and connects, I think, to many of the sentiments of people in this sort of audience interested in computational biology is ways of having simple codes. And of course, the way of one way of breaking up proteins and thinking about them is these ABCs-- the alpha helix, the beta sheet, and the coil.

Each of these can be characterized by the hydrogen bonds that hold it together. The weak bonds between the hydrogens, the nitrogens, and the oxygens. And the alpha helix, these are all kept within the helix with a repeat of 3.6 residues per turn.

In the beta sheet, they tend to have a longer, straighter chains where there are unpaired hydrogen bonds inevitably until you form enough chains to form a cyclic structure while the alpha helix is immediately helical. There are many different types of coils. It's a catch-all phrase that includes everything except alpha and beta, but a particularly well-formed type of coil has its own nomenclature and parameters is the turn, and the turn is illustrated here at the end of a beta sheet.

That beta sheet basically is going in on the lower right hand corner in the direction of the arrow. The arrows typically point from the n terminus to the c terminus just as in nucleic acids is from five prime to three prime. And here it goes in the arrow, it turns around very tightly, and goes back out again.

OK, now how can we use these basic motifs? These are the smallest meaningful units of protein three dimensional structure. How can we use these to recognize other macromolecules, other proteins and nucleic acids?

So let's connect this to the motifs of last class. We have these motifs that we could find, weight matrices for them by aligning lots of sequences. Now instead of aligning sequences, let's see what we can do by mutating both the protein part and the nucleic acid part.

And in order to do this, just as an illustration, let's say we have three zinc fingers. This is a real human and mouse DNA binding protein with three zinc fingers in a row. So this is an example of the direct repeat or tandem repeat type of symmetry. Remember, there's the direct repeat and the inverted repeat.

And in this tandem repeat, let's anchor the two ends and change the middle. Make every possible peptide sequence in the middle or randomly sample the vast space that might occur in changing a few, say, six or more key amino acids. And then we know from the three dimensional structure that it interacts mainly with the middle three nucleotides, so let's change those middle three nucleotides to every possible trinucleotide sequence and see quantitatively how much that the different sequences in the protein affect the different sequences in the nucleic acid.

So this is not going to be by staring at long sequence alignments where we're going to get the weight matrices. We can get them by actual experimental measures of the binding in vitro. And so what happens when you do this exercise, the wild type, now this-- the wild type sequence is something that may want to recognize a family of sequences.

We don't know exactly what the wild type sequence is for this particular DNA binding protein, this zinc finger from human and mouse. But the subsequence of the business end, the amino acid subsequence of that recognition alpha helix that's binding into the major groove DNA is shown in the upper left, [? RSVHLTT. ?] And the sequence that it mainly binds to-- remember this is a weight matrix. It's not a consensus sequence, this TGG. It obviously recognizes a variety of other sequences. So remember, there's about three nucleotides on either side of it that the other two zinc fingers bind to.

When you try all 64 possible trinucleotides-- remember from the genetic code, this is pure coincidence that these are triplets the same as the genetic code. This just happens to be the amount of the chunk of DNA that a zinc finger will cover. But it's not coincidence that 4 to the third is still 64, just like the genetic code.

And if you run through all possible nucleotide sequences for this wild type, you find the winner is TGG, and it has that particular binding constant. The binding constant is measured in the molarity, roughly where you get half maximal or equilibrium binding. That's 10 to the minus 9 moles per liter.

You can now mutagenize the peptide and select for peptides that bind to GCC. Remember, the flanks are kept constant. You get two peptides, both bind to the GCC. They both give matrices very similar to this with very-- these are very high affinity binding constants just like wild type. You've essentially engineered by selection a new specificity and two different ways of getting it.

If you now go for something radically different, now no GC-- the first one was high GC, the second was pure GC, and then the third one is pure AT. And you get another peptide sequence that binds to that, and you get another weight matrix.

Now remember, this weight matrix is not sequence alignment. This is binding constants where the weighting of the 64 different sequences is based on how much binding you get for each of the 64. And then for some sequences, all these different trinucleotides all result in a rather poor selection for any kind of peptide out of all the vast number of peptides you have. None of them do particularly well, and the result is a weight matrix here which has very little information content.

So this is a way. It's not the only way, but this is a way of getting a really good empirical data set, which in principle, you can combine it with similar functions on the flanking ones, and you can dial up any sequence of a nucleic acid protein interaction, at least with this class of proteins. Others are a little more problematic. But you can see how this can generate a code even if the actual detailed amino acid nucleotide interaction is not so simple.

So that those are the results of the study. Then I'll show you just a snapshot of a schematic of how the actual experiments are done. And then finally, I'll show you a little math behind how we got those apparent binding constants. Remember, in those lower binding constant means you can get the binding at a lower concentration, which means a stronger binding.

So the way you do the binding here is you take a nucleic acid array, similar to the ones we've been talking about in last class. Instead of being single stranded, ready to bind fluorescent nucleic acid, it's double stranded, ready to bind a fluorescent protein complex. The protein complex in this case is a bacteriophage, which is displaying the three zinc fingers in red. The middle one is the one in the past slide was mutagenized. And similarly, the array is combinatoric-- every possible DNA sequence that you're interested in is present.

And what you do is quantitate the fluorescence of the zinc finger protein indirectly by the binding of the covalently-attached phage to the antibodies, which are fluorescently labeled by [INAUDIBLE] fluorescence. The quantity-- how you relate the fluorescence, the more binding, the more fluorescence. But how you relate that to the binding constant we had in the previous slide is the subject of this slide number eight.

Now we call this the apparent equilibrium association constant because these experiments, just like many binding in living cells is not at equilibrium. It's a dynamic process in the cell and in vitro. There are ways that you can measure the equilibrium constants, but what this is apparent in the sense that you need to wash off the excess fluorescence in order to detect the fairly low signal that you get from the specific binding, rather than having fluorescence. You bathe it with 10 to the sixford excess of fluorescence.

And so as you're doing that wash, you're obviously not at equilibrium. In the end, you take a snapshot before you wash off everything. And so what you're measuring-- what you're basically measuring is-- attempting to measure is the equilibrium constant between protein P in the upper left here, DNA D, double stranded going to the product which is this P.D associated biomolecular product.

And this is basic physical chemistry and algebra so that-- so you rearrange that to get the association constant. That's what you want. And the fraction of DNA molecules with the protein bounds can be found from this.

It's just by definition, the fraction DNA molecules protein bound is-- the protein bound is the numerator. The brackets mean concentration in moles per liter, just like we stated. And then divide the complex by the amount of the fraction, which is the total DNA, which is the DNA present in the complex and the DNA is free. That's the D plus PD.

And then you just substitute in this definition of the association constant. The definition is product over reactants, and then you get this intermediate term. Which now you cancel out all the concentrations of these, and then you end up with the last term on the far right, which is where if you hold the protein-- if the protein concentration is known, then the whole thing is constant except for the association. So that one over the association constant is directly related to the fraction of DNA molecules with protein bound, which is directly proportional to the signal intensity and fluorescence. So this is how you get the numbers that were on that slide.

Now let's return the question that connected this talk with the last one, which was the symmetry of DNA protein interactions. We illustrated already one of these three zinc finger complexes, as illustrated on the left hand side here. The double-stranded DNA is in blue and the three zinc fingers follow along the major groove, the large groove of the DNA.

And the reason the textbook is wrong, first of all, it emphasizes the non-helical part of the zinc finger. You can barely see the helix with the background there. And also the way it loops through the DNA, if you look at this carefully in your textbook, this is actually the [? Mount ?] book, it actually interdigitates with a phosphodiester bond, basically going through the base pairs, which is not at all what happens.

Similarly, the leucine zipper, this is, again, recognition with a helix in the major groove of DNA. Here, you can see that the helix that causes the dimerization of proteins-- you can think of this as your really most elementary protein-protein interaction code. A very fundamental one that comes again and again, so-called coil-coil, two alpha helices interacting.

The direct extension of that, almost coaxial with a coil-coil [? or ?] protein-protein interaction, they go down and touch the DNA. In contrast, the textbook where it does the sharp right hand turn and in some way poorly schematized there, it goes coaxial to DNA. That's not what happens. The helices are more or less direct extensions down from the dimerization region of the protein maintaining almost perpendicular to the DNA axis.

But again, so on the left is the three tandem repeats, and on the right is a dyad axis, where the twofold 180 degrees symmetry-- think of it as rotating exactly 180 degrees. This is not a mirror. This is a rotation. Of the DNA on itself also coincides perfectly with the twofold symmetry of the protein association with itself.

These are the two major symmetry classes, and it's amazing how many nucleic acid protein interactions fall into one of these two classes-- direct repeat or inverted repeat. And that's the one you find direct and inverted repeats in nucleic acid sequences, you get a little excited. The other reason is the hairpin structures that you found in RNA that we talked about in the last classes, those also are indicated by inverted repeats.

So we now have a semi-empirical way of computing-- in a certain sense, predicting new regulatory protein DNA interactions with double-stranded DNA. Can we extend this to RNA? This is a much more complicated situation with RNA because you don't have these long perfect double helices anymore. You have these very short RNA helices that I showed in the last couple of classes.

This is transfer RNA, one of our favorite molecules here, with the anticodon at the bottom of each of the pink structures. And the amino acid acceptor three prime end of that 70-some nucleotide-- 70 to 80 nucleotide nucleic acid. So the pinks are all the tRNAs, and there are at least 20 different types of amino acid and has 20 types of-- at least 20 types of transfer RNAs and 20 types of proteins that add the amino acid onto the three prime ends of the transfer RNA.

These break themselves up into two major classes. These can be recognized at the structural level. Class 1, which is the single letter amino acid code [? CEL, ?] so forth, cysteine, glutamate, and so on. That's one class, which are structurally similar. Class 2 is structurally dissimilar to Class 1, but they are similar within the class.

Anyway, the point is that they recognize all different parts of the nucleic acid, not just the anticodon, which is the code itself, the trinucleotide code. Not just the amino acid end where you need the recognize amino acid, but various points along the transfer RNA.

If you wanted to create a new code, as these authors have, or to create hybrids between these various things, you'd have to find homology among the proteins or graph domains of recognition between each one or mutagenize particular regions that are known to interact with the nucleotides you want to contact, and that's been done. You can arrange to make a new amino acid by carving the pocket the amino acid recognizes and grafting on the appropriate nucleotides-- the appropriate amino acids would recognize, say, a stop codon.

OK, you've had some programming experience that hopefully will prepare you for the real world of interacting with input and output from various devices. The topic today is proteins, and this really is the main contact between the exquisite regulatory mechanisms, which will be the topic of the [? network ?] that we've already touched upon, but will really be the topic of some of our network analysis in the last three lectures.

Here, we need sensors to sense the environment. We need actuators to then deliver back into the environment what the cell wants to do or to interact with other cells. You have to have feedback, synchrony, so on that you can basically program the almost digital nucleic acid world inside the cell but via clearly analog inputs and outputs. So since this is the Halloween lecture and I'm masquerading as the Wolfman, we also-- I've listed some of the scariest proteins that I could think of. And we're going to talk about three of them.

One of them in the slide, which is the proteins that are actually involved in causing the symptoms that come from when you're worried about anthrax. And then we'll talk about HIV yet again, this time, polymerase mutants that cause drug resistance. And then ApoE yet again, as we have in the past, this time talking specifically about how protein structure tells us the haplotype.

So with anthrax, you start out with this simple two component-- two protein domains here. They bind to a cellular-- something on your cell surface. Hopefully not yours, but human cell surface. And then one of the domains disappears. And the remaining one now self-assembles into a seven-mer, seven-fold symmetry. Remember, we were talking about two-fold rotational symmetry for the DNA protein interaction.

This is now seven-fold rotational symmetry. That now allows lethal factor, LF, to bind-- still not inside the cell. But the whole complex gets internalized. Still, topologically, it's as if it were outside the cell when it's inside this little vesicle. It has to get through that membrane. But now the pH change that happens when this vesicle goes in the cell, part of the natural cell biological processes causes some act-- an unfortunate act where, now, the seven-mer complex of proteins does yet another conformational change and turns into this hairy beast that allows the lethal factor to get into your cell and kill it.

So you can see that when we're talking about protein three-dimensional structure, whether we're predicting it or solving it, protein is not a static object. Here, it associates with one factor. It associates seven of itself. It interacts with lethal factor. It opens up a whole new channel in the membrane, et cetera. You need to think of these as dynamic systems with many different states.

We also need to think about time scales. Many of-- the molecular mechanics we'll be talking about, the timescale of relevance is the femtosecond. You need to be able-- this is, well, two nanoseconds. So 10 to the minus 15th, 10 to the minus 9th seconds. That's atomic motion.

The turnover of an enzyme that is the time it takes to for a small molecule, say, to find and bind the enzyme, to possibly go through a catalytic step, and to dissociate as a product. That's on the order of microseconds to milliseconds. And the second range is the time that it takes the molecule to-- the drug or small molecule-- to touch the surface of the cell, maybe diffuse across the cell, and find its target.

Transcription that we talked about, all of the regulatory mechanisms of transcription last time, the rate of the constant for that process is around 50 nucleotides per second. Not entirely coincidentally, that's about the rate at which it is translated into protein. These are important numbers, because a typical gene size piece, say, after RNA splicing in higher organisms or naturally, it might be a kilobase.

So that's about a half a minute to transcribe and translate. That could be used as a timer in a circuit of these longer time frames, like cell cycle, circadian rhythm, very long time frames in ecological systems with bamboo and various pests, which can be not even yearly, and then development and aging, which can be on the order of hundreds of years, at least for humans, turtles, and whales.

So what we think proteins are good for depends on the accuracy. And the accuracy depends on the method. At the very bottom right, we have a very appealing approach, which is de novo, a priori, or ab initio prediction of secondary-- of protein three-dimensional structure from the sequence alone, which we're getting in bucketloads from the genome projects.

But unfortunately, accuracy so far-- and we'll delve into this in more detail in a moment-- is on the order of six angstroms of difference between the predicted structure and what it actually is by the more precise methods up higher on this. The y-axis here, the vertical axis, is basically sequence identity as you start doing, say, threading or comparative modeling here as you get-- instead of the ab initio or de novo prediction doesn't require any sequence similarity.

If you want to build it based on previously solved structures, you need at least 30%, but you're still very far-- say, 3 to 4 angstroms away-- from the native structure. As you get to 1 Angstrom or better in your accuracy, as you can get from NMR and X-ray crystallography, you now are in a position to study catalytic mechanism and design and improve ligands, such as drugs.

This is really where we want to be. There may be a day where we can do this all from ab initio prediction or modeling at very great distances. But for now, modeling at very short, say, 80% to 90% amino acid similarity, is important. Remember, there's a-- just like there's a variety of different protein structures.

This is just an example of a vast literature that exists where you can use some of the methods that we'll be discussing in this class on doing molecular mechanics on proteins and predicting their three-dimensional structure in complex with the various drugs.

We will contrast this, or show the interplay of the computational biology, that can be aided by actual measurements of drug binding, just as we had actual measurements of zinc finger binding to double-stranded DNA. And ways that you can discover the small molecules by a clever use of parts of it that you know bind and parts of it that you know might be variable in a chemical sense.

Now, this is what-- just as there is this dynamic competition between pathogens and their hosts, there's this similar lethal game that's played between pathogens and the pharmaceutical industry. And here, HIV, for which there are many drugs now aimed either at the protease or at the polymerase-- some of the first ones were aimed at the polymerase, and so we have a big collection. This is one of the most sequenced molecules on Earth, which is the HIV gene encoding the reverse transcriptase polymerase.

And it's been sequenced many times because as a patient takes the drugs, their population of the AIDS virus changes. And each of these little diamond-shaped substitution sites clustered around the binding site in the protein-- the binding site here indicated the substrate is in space filling, which is the triphosphate on the upper left and that the template kind of curving around on the right-hand side of that space filling bright green structure.

The protein is in red. And these little diamonds indicate substitutions, where the nomenclature is single-letter code for the wild type, the position in amino acids from the end terminus is the number, and then the third-- or the last far right letter is the new amino acids. So for example, D67N means a [? sparcade ?] at position 67 and wild type changes to an asparagine. And that causes a drug resistance in the HIV, with unfortunate consequences for the patient.

We can take-- now, making mutations in polymerases is not entirely of negative consequences. And I'm going to show you a really beautiful example where a DNA polymerase-- a very similar kind of dynamic, where the DNA polymerase, you want to change it so that it can now handle what would normally be an inhibitor of DNA polymerase, a class of nucleotides that is incapable of extending-- capable of being incorporated into the growing replicating genome, but is not capable of extending. Is a powerful inhibitor, and it's also powerful sequencing reagent, these dideoxys.

And so one of the things that was noticed as the sequences of some of these polymerases were being studied, and some of the resistance mutants, and so on, it was noticed that the complex between the nucleotide, whether it's a deoxynucleotide shown here with a 3 prime hydroxyl, which could then be extended by bringing in the next 5 phosphate.

This hydroxyl is near in space to the position on a phenylalanine or a tyrosine, position 762 of this polymerase, which can either-- if it's a tyrosine, it has an OH there. And if it's a phenylalanine, you lose the O and you just have a hydrogen there.

And when you have the phenylalanine there, there's a space that-- an appropriate space that accommodates the 3 prime hydroxyl quite well. But if you now put in a dideoxy inhibitor, you now have too much space in there, and you start trying to fill that space with other bulky molecules, like water. And basically, the binding constant, it becomes much less favorable binding when you're lacking both oxygens.

So this presented an opportunity to engineer some polymerases which had a phenylalanine there to become more accepting of the dideoxys and hence better at using disease in DNA sequencing chemistry. And this was simply by engineering-- putting in that oxygen there, by changing the phenylalanine to a tyrosine, it now made a better fit between-- you can think of it as-- you can have either oxygen, and by removing this oxygen, you replace it with an oxygen on the protein side.

I'm trying to emphasize, by a few examples here, the idea of complementary surfaces and how you can engineer them. This is a beautiful case. Now, we're talking about a single atom rather than the complementary surfaces of the nucleic acids we were talking about before. This has an 8,000-fold effect on the specificity of this polymerase, and a big impact on the Genome Project.

Now, that's how we program a particular atom to achieve an important goal. And of course, the virus has its own mechanisms for programming it typically, by random mutagenesis and selection, as we talked about in the population genetics.

But the way we program in general proteins are either transgenics, where we might overproduce the protein, or homologous recombination, which is the ultimate where we go in and if the protein is already-- if the gene encoding the protein is already present, we can change that particular nucleotide in situ in the correct place so it's properly regulated and everything. That's a great way to do it.

Point mutants are not the only way to generate conditional mutants. Many of them historically were. But there are ways that you can program, and conditional, meaning that you can regulate under what conditions the protein is expressed or not or active or not with an entire domain, or with single nucleotide polymorphisms. Now, so this is one way. This is the nucleic acid way.

Another way is by modulating the activity of the proteins from the outside with drugs or drug-like molecules and chemical kinetics. And under the subheadings for that, you can make these by combinatorial synthesis-- and we'll show an example of that. But combinatorial synthesis can be based on design principles, not just completely random. Usually are. The design principles can take into account what you know about the nature of the interaction of similar proteins.

And you can mine whatever biochemical data that you can collect for so-called quantitative structure activity relationships. This is a slightly different discipline than the detailed crystallographic and quantitative studies that we've talked about so far. Here, you're trying to basically mine through the structures of the ligand itself for the parts of the ligand that might be responsible for the activity, the binding activity or the full biological activity that you see.

So let's look at some examples of single-nucleotide polymorphisms that we've been talking about before in-- actually, this is a class that we didn't discuss before. But in previous classes. But it's related to what we've been talking about. In the case of the zinc finger, we made an altered specificity. We made new zinc fingers with bind to completely new trinucleotides. With the DNA polymerase, by changing one amino acid, we could make it now accept almost four logs better an inhibitor is very useful.

And here, many different-- many of these are enzymes, where you can not just knock out the enzyme, but actually make it recognize a new substrate, or change radically the binding constant and catalytic rate for new substrates. And I just have this long fine-print slide just to impress to you-- this is actually less than half of the list-- just how many examples. These are not that unusual. And those can be designed or naturally occurring.

Now, we're going to take the three-dimensional structure of proteins and connect it with our discussion of haplotypes and single-nucleotide polymorphisms. And you may recall that with-- one of the commonly occurring single-nucleotide polymorphisms is the ApoE4 allele. It's present in 20% of the human population.

Even though it has unfortunate consequences, we think, mainly for Alzheimer's-- it increases the risk of Alzheimer's, and probably increases cardiovascular fitness through ApoE refers to its involvement in cholesterol metabolism and transport. The ApoE3 allele is present in about 80%, and is far more common in human populations, but both of these would be considered very common alleles.

And we also mentioned that the ancestral form of this, for example, found in chimpanzees at nearly 100%, is this arginine 112, instead of what's now common in human populations was cysteine 112. And that was-- one explanation for that might be that it was physiologically-- our nutritional standards have changed. We now eat a lot more fatty things. We live long enough to get Alzheimer's.

And so maybe this was something that was-- this bad allele, E4, was good in chimpanzees that have different diets or lifespans. But the other possibility-- and I can't really distinguish between these right now-- but another one to seriously consider, not just in this case, but in cases in general, is you no longer just think about single nucleotides. You think about haplotypes.

Everything insists on that DNA strand has a chance of affecting either the expression level of the protein or insists, on the protein strand, to fall back and interact. And you can see that one of the nearest amino acids to this arginine 112, which is the main difference between ApoE4 and ApoE3. Arginine 61 is the same on the two alleles.

But you think of this as one haplotype, and in chimpanzees the haplotype, is now threonine 61. And you can think that a [? 3R ?] in chimps, or ancestrally, is not too different from [? Rcys. ?] So the different order. So it's like this compensating, complementary mutation, just like we had in the oxygens in the polymerase a couple of slides ago.

And you can think of the compensating mutations-- we had the mutual information theory for doing the theory structure. Think of complementary surfaces. When you think of single nucleotides, don't think of them as haplotype and possibly complementing constellations.

Especially, now, this brings us to the possible impact of three-dimensional structure on predicting deleterious human alleles. If we suddenly had the sequence of everyone in this room and we wanted our computer program to prioritize, which ones should I the attention to? Which deviations from the most common allele should I look at first?

Well, you might think of these things in terms of proteins. We've now gotten to the point in the course where we're talking about proteins. You need to think about the three-dimensional structure. Who's near who in the structure? You can think about binding sites. These might be indicated by-- if you know the three-dimensional structure or you know the conservation pattern in this family of proteins.

You can ask things about charge. In that last slide, we had the charge of the arginines being near a compatibly partial negative charge on the cysteines or threonines. You can have-- a disulfide is a very important thing to lose. They tend to be highly conserved.

If you introduce a proline into what would normally be an alpha helix, this is something where knowledge of the three-dimensional structure would say, oh, that proline, this a priori, without any knowledge of conservation, could be a huge change in the three-dimensional structure. And then these multi-sequence profiles are a good way of looking at the conservation. That's a way of prioritizing single-nucleotide polymorphisms that might have impact on pharmacogenomics or disease in general.

Now, as we integrate that with the chemical diversity that we can create-- that's going to be the topic for the next few slides, is how do we create chemical diversity? And I'm going to introduce this, the idea of chemical diversity, in a way I hope nicely connects to where we've been with RNA arrays. RNA arrays, and the double-stranded RNA array that we used earlier in class today, can be generated in a commentorial sense. You can make an exhaustive set.

Now, typically, those were made where, spatially, they were isolated. Each different nucleic acid, oligonucleotide, is present in a different place identifiable to the computer by its coordinates on an array. But you can also make them in a big mixture and use them as a mixture and do selection on them, as we did with the phage display. Or you can make them as a mixture of solid phase particles and then separate the [INAUDIBLE] phase particles out in some manner.

Solid phase comes up again and again in arrays. It's very obvious why you have a solid phase. You want to be able to address it by its positions in x and y on the array. But the other reason-- technical reasons are, it's a fantastic way of getting purification of your products simply by washing rather than doing complicated purification procedures. And it allows you to, in the case of beads, there now-- you can think of it as an ultimate and flexible array. You can move the beads around and put them in new arrays, and identify them later.

Anyway, so we're going to introduce the general way of making either-- complex chemicals, whether they're linear polymers like proteins or nucleic acids, or much more tighter and small molecules. But they have similar concepts that you need. There's the solid phase that I already talked about. There's the idea of protecting groups, and the protecting groups are protecting against a reactive group.

So the highly reactive group here is the phosphoramidite, which is this phosphate nitrogen bond. This is capable of reacting with just about any nitrogen or oxygen, such as this-- eventually, once you do protect this oxygen at the 5 prime position, these two oxygens are 5 and 3, just like the ones we've been talking about all along. This is the chemical synthesis version of the polymerase that we've been talking about. So you have these reactive groups and the protecting groups. And those are the major concepts.

Now, let's go through. This is-- the topic here is proteins. And we'll talk more about protein synthesis as part of quantitation next time, and as part of networks in the last three lectures.

But here's a completely synthetic way of getting short peptides, either by directly synthesizing the peptides or synthesizing nucleic acid that encodes that peptide or interferes with the production of that peptide. And you can think of these as drug-like molecules. These are naturally related-- they can be analogs of nucleic acids and proteins, not just straight ones. And we'll talk about opportunities for making these analogs.

And so by making analogs of known proteins or nucleic acids, you kind of have a more immediate connection between the thing that your computer instructed the synthesizer to make and your targets. Well, if you make a random chemical, you don't necessarily know what your target is. But we'll talk about ways of making slightly less random chemicals. But this is one way of making a direct connection.

And the process is cyclic in the sense that each cycle, you return, and the polymer gets a little bit longer. You start with one monomer on a solid phase, shown by these little hexagons on the far right side of the slide. And you add-- you remove the protecting group on the immobilized polymer, one protecting group. And then you bring in this reactive group, otherwise protected.

And there's really one major product that you expect. You wash off all of the excess. You now have one longer. You deprotect. This DMT group is removed. And you go back up and cycle again. There may be additional steps, such as oxidation, which will stabilize the new bond that you've made. Or you can have a capping step that can soak up any excess that was left over.

But in general, after you're done with all of these cycles, then there'll be a step where you remove the protecting groups altogether and remove the polymer from the solid phase if you so choose. Or you leave it there, if you have an array. These are other examples of protecting groups. These are now on the bases. Some bases don't need protection, like thiamines. If you have an exocyclic amine, then you typically need a benzyl or an isobutyryl group in blue here, are the protecting groups.

I said there's an opportunity here for modifying the nucleotides or oligopeptides or other chemicals to make them so that they're related to, but not identical in every property, to normal constituents of your body or of a bacterial cell that you're aiming at. And why would you want to make derivatives? Why not make the exact thing?

And the reason you want to make derivatives would be, for example, to increase their stability, or decrease their stability, or make them bind more faster or more irreversibly. And examples are, in the previous slide, you can make modified bases. And in slide 29 here you can change the backbone itself.

You can change the ribosomes so that they have bulky groups that prevent the nucleases from getting in, or they can exchange oxygen for sulfurs or hydrogens in order to make the phosphodiester backbone itself, which is where nucleus is cleaved, a less attractive, less energetically favorable substrate.

So now, those are chemical processes that, in a certain sense, mimic the normal polymerases and ribosomes in the cell. And there are analogous processes to generate chemical diversity on smaller molecules. And then there are, analogous to that, biological mechanisms by which you can make small molecule diversity which are less cyclic than processes we just talked about.

These are more a set of ordered reactions that has a conceptual repeat, but in a sense, you can think of it as a linear program that you go from the beginning to the end. And that's to make these polyketides, which are shown on the right-hand side of the slide. A large class of pharmaceuticals, including most of the antibiotics, are made by a fairly small set of organisms, such as streptomyces in certain plants.

And this process by which it's made, which will be illustrated in more detail in the next slide, is very akin to fatty acid synthesis. Fatty acids are long hydrocarbon chains, which you can think of, adding each two carbons on that fatty acid is a process akin to the one for making these polyketide drug-like molecules.

Another way of biologically making a very compact structure, which actually uses ribosomes, but it uses them to make very tight short peptides and a precursor that then folds with lots of disulfides to make this small and highly cross-linked.

Now, the fact that these cone snails have gone have gone to the trouble of making hundreds of these different very small peptides that have these properties tells you something about what it is that you want-- that drug-like molecules have in common. And that is, they're small, so they diffuse quickly and get to their site. And so you can have large amounts of them in a small space. So you can manufacture lots of them.

And then they have to be highly cross-linked in order to maintain the rigidity. Because they're small, they have less surface area to bind to their binding pocket. So to compensate for that, they have to have a lot of rigidity. And the thing that you lock in rigidly has to be the correct structure. It does you no good to have a rigid structure that isn't really perfectly complementary to that surface.

And the third source of biological diversity is one you're probably more familiar with, which is the immune receptors, the B and T cell receptors, the antibodies, and cell-mediated immunity. And these use recombination machinery to program various combinations of nucleic acid motifs that encode protein motifs.

And as they do that recombination, they have further diversification that occurs due to a template-independent polymerase, [? thermotransferase, ?] which will extend a few nucleotides of completely random nucleic acid sequence that basically incredibly accelerates the rate of mutagenesis, basically generating sequence de novo. This is one of the examples in biology where you generate sequence de novo. And I think that's very apropos of this combinatorial topic.

Now, this is a beautiful example from the previous slide. Those polyketides on the far right now are the star of this slide. And here, in a certain sense, nature has-- and now scientists have-- engineered protein modules to make this linear sequence of events.

You can think of, here, you're using a linear set of protein domains to program very complicated series of chemical reactions, the same way the linear sequence of messenger RNA tells the ribosome to make a series of additions in the protein. The ribosome catalytic cycle is a cycle, while this is more a linear tape, a linear series of events.

The proteins themselves, these little arrow-shaped things with boxes in them along the top, labeled Module 1 through 6, those proteins are, of course, made on ribosomes. But then they act kind of like the solid phase synthesis, where the acyl carrier protein, ACP in the box, binds to the first monomer, and it starts transferring it from protein to protein along this multi-domain huge protein. And there's actually three proteins in a row here.

And each of the steps are taken in order along the protein, and they involve things like a synthase step, where you bring in another monomer, or a keto reductase step, KR, where you'll bring in-- where you'll reduce one of the double bonds, or an acyltransferase step, AT.

But you can see each of these has a substrate specificity. And by changing the order of substrate specificities, you can build up a huge combinatorial collection here in microbial communities, and also in the laboratory.

Now, protein interaction. We're just beginning to talk seriously about protein interaction assays. I think many of you either are or will be more and more familiar with the protein interaction assays. In next class, we'll talk about ways of getting direct information on protein through cross-linking and mass spectrometry.

Another way is indirectly setting up these reporter assays, where you take advantage of the binding properties of known proteins to analyze two unknown proteins. And so a known protein might be, [INAUDIBLE], which binds to DNA, and B42, which activates transcription of something for which you have a good visual assay, like [? URO3, ?] life and death.

And by taking these two knowns, let's say, in B42, which have known properties-- but they only exert their properties when they're brought together, and they're only brought together if the unknown proteins or partially known proteins to which they're bound interact with each other.

And so this is a so-called two-hybrid assay and variations on it. And I think we mentioned one where you can characterize nucleic acid-protein interactions with one hybrid assay. And here, you can inhibit this interaction between the two knowns-- sorry, unknowns or partially known molecules in blue here.

Here's a TGF beta, a growth factor, and a binding protein. You can inhibit that with a particular small molecule or a collection of small molecules, which can be introduced from one of these combinatorial syntheses into an array of these cellular assays. And you get this information about-- you can either collect a big data set of proteins that interact from a proteomic scale experiment, or from molecules that inhibit one or more of those interactions.

This is a source of information, which is intrinsically computational in the sense-- well, in the sense that there's a large amount of it. You can model the three-dimensional structure of this interaction if you have sufficient data to do that. And you can model the impact of the small molecules in a structure activity [? since. ?]

Now, if you look at the top right-hand part of slide 34 here, you can see this huge diversity of all of these different colored shapes. And if you wanted to use these in a combinatorial assay, you'd connect them in every pairwise combination and try them against your target by some bioassay.

However, if that library is too large either to make or to screen-- typically, to screen-- then what you can do is, you can study a part of the molecule and see-- and then take the subset of the diversity that can bind, and then take that subset, characterize it, and now make only the pairwise combinations of the two half-molecules.

And generally speaking, if the geometry is fairly rigid, then the binding constant will get-- will be roughly the product of the two binding constants. So if each one has a very low binding constant, then it will be roughly the square of that. And you get some point of diminishing returns, eventually. So this is an example of a strategy where you use a little bit of prior knowledge, which can be empirical or it could be purely computational, about how to limit your library and make interesting combinations.

Now, we've been talking mainly about the kind of chemical diversity we can get that's aimed at the ligands that combine to target proteins. But now, we want to talk about the source of information about those target proteins itself, which is another genomics project, structural genomics. And typically, we want to select targets for binding drugs, or select targets for solving the structures of proteins in order to look at their ligands in more detail.

And how can we do this? How can we decide which targets are high on our list to go for next? We have hundreds of proteins for which we have three-dimensional structures, and from some of them, we have information on what ligands they bind. But these are other criteria that are sometimes used in the field for target selection.

If they are homologous to previous interesting targets, then that puts them high on the short list. If they are conserved-- and we've talked about how important conservation is from time to time-- then that might be an approach. If they're conserved and you knock it out, then you might expect that to be lethal. And that might make it a good target for an anti-bacterial.

If you want to limit the action of your therapy to the surface in order to, say, reduce the cross-reaction with internal molecules, you can sometimes restrict yourself to the surface-acceptable proteins. And in fact, a very large class of drugs is aimed at surface-accessible membrane proteins. And so very often, those are prioritized high.

There are also-- the surface-acceptable proteins are important if you're talking about vaccines, which is an increasingly important-- or diagnostics that are non-invasive, or at least not going to the interior of the cell. And there are ways, say, with microarrays, that you can ask which genes are differentially expressed in the disease state. And that causes high prioritization.

Now, once you have that prioritization, which comes from, say, genome sequencing and some of those other facts, then we would like to have-- you've got your target. You've got your genome sequence-- gene sequence for that target. How, then, do you get the three-dimensional structure that helps you design drugs or improve the drugs that you have?

Well, one very attractive approach, given a protein sequence which might get from their deluge of genome sequences, the practical approach might be to start with this genome sequence-- this gene sequence which is 99.99% accurate, and try to predict the three dimensional structure of the protein and its ligand specificity.

And if you walk through these, these are kind of ballpark estimates, some of them better than others. To get from the sequence to exons-- we talked about this before-- might be an 80%. Remember, these numbers really should be false positives, false negatives, so on. But this ballpark, 80%, getting to exons, exons to genes. If you aren't privileged enough to have the cDNA, then this is an error-prone process with maybe a 30% success rate.

Once you have genes knowing it's regulation, whether it's on or off in the particular cell types you're interested in, is very difficult right now. Knowing the motifs is barely a start on getting the full regulation. So I would say 10% or less. Once you have a regulated gene, getting the protein sequence is easy. That's the genetic code.

Getting from the sequence to secondary structure is easy in the context of some of these other things, but still, the accuracy is only around 77%. I have next to this the reference is [? cast, ?] which is something that is a competition for computational assessment of three-dimensional structure prediction for proteins that's been held since, I think, '94, and the next one is coming up in a couple of months.

And this is kind of the big race or bake-off between the different methods. Very exciting. But unfortunately, over decades, it's still hovering around 77% for secondary structure and about 25% for ab initio three-dimensional structure.

Then even if you have the three-dimensional structure at adequate accuracy, getting the ligand specificity is problematic, and it depends on the ligand. If it's DNA, it's a good case. If it's a small molecule, it can be as low as 10% or worse. Now, since each of these are fairly independent of estimates, you can get an approximate overall accuracy which is a product of all of those, which is dismally small at 0.0005.

And so it behooves one to use as many extra-experimental data as one can, or improve the algorithms that are weakest in this journey from genome sequence to ligand specificity. We'll pick up this thread right after a break and carry on to actually how we get the three-dimensional structure, whether it's predictive or experimental, and the computational tasks there. Thank you. Take a break.