The following content is provided by MIT OpenCourseWare under a Creative Commons license. Additional information about our license, and MIT OpenCourseWare in general, is available at MIT.edu.

PROFESSOR: OK. Welcome back to the second half. We'll pick up on this somewhat pessimistic note of how one goes from a very accurate sequence to a probably less than accurate speculative league and specificity, to actually how we actually do get a 3D structure league and specificity, and some of the powerful methods that are shared by the computational tools.

One of the things we want to do is we want to, if possible, find a homologue. It may be a distant homologue or a very close one. If it's very close, as we saw from that previous slide about the highest accuracy coming from homologues that are 80% or 90%. As we get further and further away, remember we had to-- first, you could use exact matches. And then you had to use dynamic programming, and then hidden Markov models.

And finally, at the bottom of this Slide 37, we go to-- we resort to threading, where one of the structures is searched through a database of three-dimensional structures with your favorite sequence. And you not only search through the database of structures, but through each structure, you think of every way you could thread your sequence into that complicated 3D structure.

And some of the threading positions with insertions, deletions, and offsets will be better than others. They will fit that three-dimensional structure. Some will cause clashes, where the three-dimensional structure you're searching through and the particular thread position you've got causes two bulky groups to lie on top of each other in space in a way that's hard to relieve the stress.

And so, that threading is probably the ultimate in getting very distant relationships. It is limited by the fact that it really only works if one of the two proteins has a three-dimensional structure. But you can search the database of sequences against a database of 3D structures.

The antidote to the limitation of having not enough three-dimensional structures is to launch a project, like the Genome Project, to get all of the three-dimensional structures. Now, is that finite? Well, certainly for a particular genome, that number is less than or equal to the number of proteins in that proteome for that particular organism.

But if we look at organisms as a whole, where we don't even really even know how many organisms there are on Earth, then that number may be larger. But some people estimate that it's less than 10,000. Basic folds, where a fold is the scaffolding-- which, once you have that fold, whether it's a certain number of alpha helices, betas and turns in a particular order, in a particular geometry.

Once you have that, and you have any amino acid sequences within a 35% amino acid identity-- and you'll see, as this last section of this lecture goes along, why it's around 35%. The goal is to get, saturate the three-dimensional structure space, so that every three-- every sequence is within at least 35% of one of those structures.

This is somewhat conjectural, unlike the Human Genome Project, where we knew we had 3 billion bases to sequence. Here, we hope that we have 10,000 basic folds. And we hope that 35% amino acid sequence identity will be enough to do homology modeling. OK? It's not. It's not currently.

And the criteria for this has to be prioritized in some way. And remember, we had prioritization for drug targets in the previous slide. And here, prioritization for structural genomics is similar. But in addition, you want to have-- you want them to represent the largest family you can get, but not have it previously solved.

And for some reason, they're excluding non-- or they're excluding transmembrane proteins. Now, this is a very important class, as we'll see in the next slide. Because the goal stated up at the top of this, Slide 38, of assigning functions and interpreting disease related polymorphisms and drug targets and so on certainly apply to membrane proteins as well.

And there are reasons that we've already alluded to for looking specific for programming cells via membrane proteins. This is where cell-cell interactions occur. This is where adhesion motility immune recognition occurs. These all occur without getting inside of the cell. This is a major class of drug targets.

And furthermore, it's not like this is an impossible class of proteins to solve. Actually, the three-dimensional structural databases-- more about that in a moment-- are filled with these things. They're certainly-- are the most underrepresented class, but there are plenty of examples. And there are two major classes.

One of them is soluble fragments of fibrous or membrane protein. Some of the fibrous proteins are excluded as well. Here you'll use a protease to cleave off, possibly, a tiny piece that makes it insoluble, maybe a little anchor into the membrane. And all the rest of it now behaves like a soluble protein. And we know how to solve soluble proteins.

Other times, the other class is integral membrane proteins, where they go like this one bacteria on the right-hand side of the image, you can see the red alpha helices go back and forth across the membrane where the gray lipids. And this is-- there's no way you could clip off a little piece of this and have a major fraction of it left to solve.

But this was solved in the membrane. And you can see the little blue water molecules going through that channel in this. That channel is responsible for a proton pumping, which can be part of the ATP production process. But there are many other classes-- redox proteins, toxins, ion channels, photosynthesis and phototransduction, and so on. The G protein coupled-receptor class is a particularly important drug target. ABC proteins and transporters have also been solved.

So given that this is an underrepresented class, and given that the structural genomics project will not necessarily target these in a rapid manner, what is the current state of affairs for computational prediction of the transmembrane regions of proteins? And actually, I would say the prospects here are fairly favorable compared to some of the least favorable ones in that slide, that very pessimistic slide a few back.

Here, you can get, as indicated in this JMP paper, transmembrane helices identified-- and there can also be transmembrane beta, as well-- but helices identify with accuracy greater than 99%. And remember, this is basically saying that you correctly identified those.

And then you also have false predictions. There'll be a number of peptide regions of known proteins, which are incorrectly predicted to be in transmembrane. And this is a tolerable false prediction rate of 17% to 43%, given a set of soluble proteins as a negative learning set.

Now, that's-- just merely knowing that a particular segment of protein is transmembrane is a big step in terms of identifying its function. But to get the further functional characterization, we need things like ligand binding, which we've already addressed. And if you look at some of these quotes, you can see that a lot of the emphasis is on display and cataloging and a hopeful expectation that we'll be able to move from rough three-dimensional structures to ligand-binding specificity.

But where do the three-dimensional structures that we do have, do believe, that do tell us about the exact geometry of ligand binding come from? Where do they come from? And how do we compute on them? How do we read them? How do our computer programs read them? Well, this is a typical file of a three-dimensional structure. This happens to be one that we will show at the-- the three-dimensional structure at the very end of this talk.

It is the human estrogen receptor. And you can see, the first line is that it is a complex between a protein and a DNA molecule. And the molecules, the estrogen receptor. The third line down is the resolution. This is a technical description of the X-ray diffraction pattern. It gives you an upper limit to how precise it is.

It's going to be more precise than 2.4 angstroms, depending on how much statistical oversampling you have, and how good your computer program is enforcing their chemical constraints. A typical precision for a 2.4 angstrom protein structure might be on the order of 0.3 angstroms, maybe eight times better than the nominal resolution.

But that's an important number, unambiguously determined in the process of collecting it. So when you look at the literature, look at this number. And look at the next number down, which is the R value, which is not a measure of resolution, but a measure of goodness of fit between the model.

The model is the X, Y, Z-coordinates of your atoms. It's the goodness of fit between the model and the data. And we'll have a slide coming up soon of how this R value is calculated. OK. The next line down begins the sequence. And if you have a multi-chain sequence, here I've cut-- I've cut out some lines for-- there's many lines of sequence for the protein in this three-letter code, and main lines of sequence for the nucleic acid in this one-letter code.

And then, additional chemical parts of the structure. Remember, the structure is complicated. It's not just protein. Here it has nucleic acid, has zinc. It has water molecules, sometimes various other things. Each of these molecules, if you can find it in the structure, you will determine the X, Y and Z-coordinates.

So the next one tells you the secondary structure. Remember, that there's three basic types-- alpha helices, beta sheets, and coils. And these are described-- and again, I'm just showing you one line example of each. There's a long list for each of these, where they've been identified by either manually or a computational automation from the structure. And these can be useful as a summary of the structure.

And then, here's the real meat of the structure. The lines that begin with the word atom, this is a position of the nitrogen atom number one in methionine, which is the amino acid number one in the A chain. So it met A. The A-chain happens to be the protein chain. And then, following that is the residue number one.

And then, XYZ coordinates, roughly 50, 24, 79. Then a scale factor one, which is almost always one. And then a B factor, 60, which is representative of how far from that XYZ value can it deviate? That's a square deviation term. And it absorbs the thermal motion of that atom and various structural defects. So it gives you some idea of the disorder of that atom.

And then you have the last couple of records have to do, what atom is connected to what atom in the structure? In a certain sense, those can often be inferred just by the distance between atoms in the structure. Now on the far right hand side, it's just the record number and the shorthand for the structure, which is 1/8 CQ. 1/8 CQ refers to the human chicken receptor.

So that's a very dry-- that's the way that it appears when you download it from the database PDB or RCSV. Then when you display it, while you're solving it, if you're NMR or X-ray crystallographer or possibly display it from the databases, and the two different cultures in [INAUDIBLE] on the left tends to describe their structures as multiple chain tracings because they want to either express their uncertainty of the structure, or they want to brag about how they know something about the dynamics.

Whatever, you have multiple chains which overlap here in different colors indicating some of the different uncertainty or dynamics of each major atom. Sometimes you'll show all the atoms as one on the right, or you'll just show the major atoms like the carbon alphas which is the center of each amino acid as you go along on the left. On the right is the way that X-ray crystallographer might show the fit to the data.

So the model is a stick figure connecting the atoms as circles. And then the mesh work is the electron density, which you can observe once you have all the X-ray data and the model, or you can calculate it once you have the model. The model plus the known physics of scattering of each of the-- known physics of the electron density of each of the atoms, you can calculate electron density.

Now you can compare the calculated electron density with the observed. Or you can compare the calculated scattering with the observed. Typically, it's done in the scattering, which is the 4E transform of the electron density. And that's all this is. The electron density is indicated by RHO here in the middle of this formula. And the 4E transform is just this integral of RHO over the phasing information, the phasing of the light waves.

These waves, just like a wave on an ocean has a phase. Whether it's up or down in the trough, how much. And so the product of the RHO electron density, which is a function of X, Y, and Z, all three coordinates, its is summed by these integrals-- It's a continuous function-- from 0 to 1. 0 to 1, and X, Y, and Z. This is Y01, it's because this is a repeating structure. It has a little rectangular cube of space around it, which repeats.

And so all you really need to do to calculate the entire electron density is to think about this little cube, which goes from 0 to 1 in those arbitrary units. But now that's how you can get from one space to another. From the electron density to the scattering that you actually observe when you shine x-ray light upon a repeating crystal structure. But now you want to adjust the model. Adjust those atoms in the previous slide so that you can maximize the fit.

That is to say, minimize the difference between the observed scattering F0 and the calculated scattering, FC. Because you know the scattering of each atom and you know that you're trying to determine the position of each atom. The position of the atom might be a parameter, P, which you adjust a small bit at a time.

And that and that change can be approximated by what's called a Paler Series Expansion. Here, we're just taking the first term, which involves first derivatives, all sub-sequence terms involve second derivatives and higher. And those are close enough to 0 for this work that you drop them. And this basically says that if we're going to adjust this parameter, we can get a feeling for how fast to adjust it, which is based on the sensitivity of the scattering, the F. The derivative of F calculated with respect to each parameter. The parameters would be X, Y, and Z-coordinates, or they could be some kind of rotational parameters.

So this is to give you a flavor for how it's actually done. This is how you actually get from the scattering off of a crystal. The crystal has the advantage. In principle, you can do a scattering experiment from single molecules. But single molecules, the signal is too weak. And it's swamped out by the noise of random other photon events.

So by having a large number of them in ordered array, they all cohere and they basically do your statistics for you. They integrate and you get the value of the statistics of billions of molecules without having to observe each of the billions of molecules and then do the computation in the presence of a huge noise. So that's what the crystal is all about.

NMR also requires billions of molecules. And so they both have the big demand of requiring large amounts of pure molecules. And that's one of the reasons that membrane proteins have been harder to get at. It's harder to get large amounts of pure membrane pieces. Now these two methods NMR, which I won't describe, and extra crystallography that I barely described, share with the abinitio methods of protein structure prediction certain key computational components.

And these are embedded in a combined system, which does crystallography, NMR, and some of these molecular mechanics that keeps the structure. You can imagine that if the structure started blowing off, atoms going in weird directions, it could still minimize the function if you have a local minimum. But if you hold the chemistry intact and satisfy what you know about molecule mechanics of chemicals, then you actually can fit a structure from further away.

Now here's the ARC factor I said we will come back to. As you do this refinement, as you adjust the positions of each of the atoms, as your computer adjusts the position of each of these atoms, you compare the scattering of the observed F0 with the FC. These are in absolute values because actually, the scattering results in loss of phase information.

So the actual things you measure are absolute values. And then you take the absolute value of the difference in order to make sure you sum a positive number. And then you normalize this, as we did before, to put it on a recognizable standard scale where 0.4 means you have a very crude structure-- if you see this in the literature you don't believe it. If it's less than 0.25, which the last structure was, then you believe it as pretty close to done.

This is very analogous to correlation coefficient. Remember, we had a linear correlation coefficient between two functions, here would be observed and calculated. If they correlate well, then you're getting close to done. Correlating well is better than 0.7, in this case. And one way of reporting the similarities between two structures, this is not a goodness of fit between model and data, this is a goodness of fit between two models. Atom by atom you go through and you measure the distance between them.

And that root mean squared deviation, of all the distances over all the atoms, or all the key atoms, core atoms, carbon alphas-- or maybe even a smaller core than that-- allows you to quote a root mean square deviation, which has some meaning independent of how many atoms you have and what proteins you're looking at. Each of these is try to put this on a common scale so you can compare from structure to structure. Now if we're going to do molecular mechanics, which is common to the computational empirical methods and the computational sequence based methods, we need to talk about the side chains of the proteins.

We've mainly been talking about the backbones. And just a refresher, this is from the geniculate code. Again, the blue or the positively charge and the negative charge and so forth. They have a chirality. It matters whether you're talking about L-amino acids or D-amino acids. The way you remember-- this is just a mnemonic for remembering it-- is that when the hydrogens point towards you, going clockwise, it goes C0 carbonyl R, this is a side chain in corn.

And some of the 19 of these amino acids have a chirality there. Glycine does not because it has two hydrogens instead of an R, it has a hydrogen. And two of the amino acids actually have two centers of chiral asymmetry. Threonine, which has this side chain, and Isoleucine. They have the carbon alpha and the carbon beta are both asymmetric.

And one of the very earliest exercises was done when the very first models of proteins were looked at, little peptides. Can do this by hand, with some very simple crude models. And you can go through systematically, there are three bonds along a peptide, long peptide. And these are the peptide bond itself, which connects one amino acid to the next one. And that tends to be pretty rigid.

It has a partial double stranded bond and it tends to be a trans-configuration, 180 degrees. This is the rotation around the bonds, not the bond angle, but the rotation around the bonds. And there are two other bonds that are not so rigid. So these are free, but they're constrained by the clashes that occur that when you rotate around the bond, the side chains will clash with other parts of the protein.

And so Ramachandran and colleagues went through systematically, all the possible Phi and PSI angles, these are these two free bonds. And this is shown here ranging over the full range of Phi and PSI on the horizontal vertical axis. And you get these little orange regions where even with very bulky chain groups, which would occur you get these allowed regions. And these two allowed regions happened to coincide with two of the most popular motifs you find in proteins, which are the beta sheet and the alpha helix.

There are other things, such as the 310 helix and various other structures that turn up. But those are by far the two most common. And the yellow shows how they get extended when you have smaller side chains that allow more parts of the conformation spaces as it's called, to be inspected. Now that's a very crude thing that you can do with very simple stick figures.

But as you get to more detailed analysis, the ultimate application of all we know about physics, if we could compute them would be quantum electrodynamics. This is way out of range for any molecule of the size that we're interested in. And then as you go down this list, you get more and more precise programs until you get down to something which is barely computationally feasible for things the size of proteins. And a great approximation of all the quantum approximations above it.

Every one of these is an approximation, but each one as you go down, gets more and more approximate. And the main thing that's missing from molecular mechanics that's present in the next step up is the polarization of electrons. In other words, in micro mechanics, you assume the electron clouds are basically spherical. And this is a huge loss, but it still is computationally very demanding. So you don't get that asymmetric polarization that you get in hydrogen bonds and many other dipoles.

So this is really-- this is basic physics. You can see the first line, forced equals mass times acceleration. Basic Newton's law. And Newton also introduced the calculus to us. So he would be very comfortable with the next line, which is that force can be redefined as the first derivative of energy with respect to position or radius.

And then mass is just mass. And we're introducing the subscript I, for the atomic. For each atom gets its own Newton's law. And then acceleration is just a second derivative of position with respect to time. Now what kind of time constants are we talking about here? This is the femtosecond range for atomic motion to the minus 15 seconds. And as you step through, you update this kinetic procedure, you can do it in half time intervals, updating velocity and position every femtosecond or half femtosecond.

So now what's this energy term? This is what I alluded to in the previous slide as being very approximate. And semi-empirical, it is based on experiments not entirely from first principles or not even from the quantum approximations. You have, say, spectroscopic analyzes that will show that the spring-like motion that two atoms can have when they're connected by a bond has a kind of a Hooke's law type of spring motion.

And that's the energy of the bond length, EB in this sum of all the E's, slide 52. And E theta is the angle that you have as a bond angle bends. And that's the spring-like force. And then omega is this kind of torsion angle that we've been talking about in the phi PSI plot, Ramachandran plot just before. Van der Waals is the non-bonded contact, which can be either positive or negative. Actually, it should show down at the very bottom of the slide is that there is a repulsive force, which is related to the R to the 12th power. And an attractive force, which is R to the sixth power.

So as you get closer, it starts to get attracted until you get this hard sphere repulsion as you get a little bit closer. Electrostatic interactions are the longest range effects. All these covalent bonds, B and theta and omega, are short range. Van der Waals are short range. Electrostatic is slightly longer range because it's a 1 over R, or R is the distance between the two atoms. And those are the main terms that enter into all molecular mechanics, whether they're used in crystallography or whether they used in abinitio.

Now this is the state of the art for abinitio. Just the very most recent CAS competition resulted in a very clear winner by some criteria, at least. The Baker Lab here, the URL is down here, is the number of standard deviations away for the mean in terms of the score for the number of correct predictions, here out at around 30 where the mean is close to zero. And even with this huge advance for the field in prediction, still this is a typical RMS standard deviation between the real structure, which was kept hidden from sight from all these competitors-- until it was known, but not to the competitors-- and then revealed. And the RMS deviation was 6, or 4, or 5 in that range depending on the structure and whether you include all the atoms or just the core ones.

And this is not adequate, as we saw in that slide actually from the same group earlier on. Another way of looking at this is now-- those were predictive structures-- these are now observed structures. The purple is comparing two structures, both of which were done by X-ray crystallography. And along the red axis here is sequence identity, ranging from 0 to close to 100%, say 96 plus percent. And the green axis is the RMS root mean square mediation between structure one and structure two.

And you can see that-- think of this purple curve as starting in the lower right, where you have very high sequence identity and less than 1 angstrom written in deviation. That means that when you solve two proteins that are very similar in sequence, you will get very similar structures. That's good. That bodes well for structural modeling, although that is not structural modeling, I mean, not homology modeling. Then as you go down in sequence identity, the purple curve starts to slope up and up until it starts curving up towards 2.74 and beyond. it gets harder and harder to do these structural alignments.

And so 4 angstroms is the sort you would get from homology modeling at less than 20% or 30% sequence identity. And this is what I said earlier, why we're trying to get enough proteins populating. This is all known proteins here being compared, when enough populating it. So you never have to go below 35% into this Twilight Zone, where you really can't make good-- you don't find good RMS deviations between two known crystal structures.

Now as we do protein dynamics using the molecular mechanics approximation we talked about, these can be applied not only to predict a static structure or a series of steps in a protein process, but the dynamics of folding from a completely unfolded protein as it might be coming off the ribosome. And this is something for which there are relatively few experimental methods. And so this is clearly a valuable contribution, but there's a problem with doing a theoretical calculation that's hard to empirically verify. But in any case, to do one of the larger tasks and IBM and others are sinking significant resources and infrastructure of this.

But doing your femtosecond time scale over a one microsecond simulation, you can easily do the math, that's 10 to the minus 6, divided by 10 to the -15 is about 10 to the ninth such steps, each of which involves this big calculation that we just went through all the energy terms log. But that's been done for this. And you can see the blue and the red represent the calculated and the observed structure at one point in the dynamic simulation.

When you have a protein three dimensional structure, you can try to dock it with small molecules. This could be easier, in principle, because you can keep both the small molecule and the protein relatively rigid as you dock them. There has to be some flexibility hence the name, flex, for one of these programs. And overall, the results are intriguing enough that you might want to use it as an alternative in the few cases where you have the three dimensional structure of a protein, but for some reason you can't solve the three dimensional structure of the complex.

But you must remember that actually, even though we cited that the solving of protein structure might be $100,000, solving a complex once you have the protein structure is actually considerably less than that. But in any case, this is encouraging where you have in the order of 0.25 to 1.84 as a root mean squared deviation between the predicted and the experimental binding modes of small molecule. You can imagine that to be off by 1.8 angstroms, it must be docking in roughly the right pocket, but maybe at the wrong angle or yeah, maybe slightly off.

So the last topic is the issue of cross talk. As we talk about protein three dimensional structures, we try to find homologs. And we often find homologs within an organism, pair logs. And these pair logs and alternative splice forms of a protein are potential toxic side reactions of a particular drug. And you can see that many of these drugs are aimed at family members. For example, a top two are part of the steroid binding family, which we have already introduced once and it will be in an upcoming slide.

And when you consider that these proteins, that particular class of proteins interacts both with a small molecule, which is either a natural or artificial steroid or thyroid like which is a steroid like compound, and it binds to a target nucleic acid. And both the nucleic acid and the small molecule have potential for crosstalk. And here is the nucleic acid part of the story. And in the next slide will show the small molecule part of the story.

But the nucleic acid part, you have two protein domains similar to one another. This is another example of the symmetry that we started this talk and ended last talk with. The symmetry here is, you have these two that can be direct or inverted repeats, separated by little spacers here. So the DNA is in yellow and the little spacers are in the gray and CPK colors. And the protein domains are in green and white, where the green and white are structurally similar to one another. It's hard for you appreciate them going around like that.

This is to emphasize that the direct or inverted repeat here. Now that's the DNA interaction. And this is the ligand binding. You can see the estradiol is the small, yellow ligand. And the tamoxifen, which is the larger ligand. This is something that's important in treating breast cancer that might be responsive to estrogen binding drugs.

So this is the part of the protein that has two parts, or three parts here. That's the binding domain. The little red thing is an activator peptide and then there's the DNA binding component. Now what are the crosstalk we have here? You can see that these wide variety of different steroid-like protein binding domains, they bind a vitamin D3, retinoic acids such as those that occur in developmental processes and in vision. Thyroid, which regulates our metabolism. And estrogen, testosterone, and so forth.

All of these things have fairly similar small molecule binding sites. And the DNA sequences they bind are these half sites, which are very closely conserved in all the members of this big family. And one of the main differences is the distance between these can vary. And the distance here is indicated on the far left hand lower left here. dR3, means direct repeat with three nucleotides in between those two halves sides. IR0 means an inverted repeat with 0 nucleotides between the half sites. The R15 is direct repeat with 15 nucleotides, and so on.

And you see each family member has a distinct ligand and nucleic acid. Although, there's a lot of similarity of the ligands, a lot of similarity of the nucleic acids. How do we-- last line of this slide 61-- target one member of this protein family or other protein families? In some cases, you will have complete artistic control, not only on the small molecule, but of the protein itself. If you have a small molecule that looks like ATP, you can inhibit all sorts of ATP binding proteins.

If you're lucky, you can inhibit a specific class of ATP binding proteins. But knocking out a particular member of a class is hard. And you can see here on the right hand side, these three chemical structures. The adenine part are these five and six member fused rings. And attach to them are the side chains. The first one, all black structure, is a known inhibitor of protein kinases in general. And the red additions are how to make that a little bit bulkier so that it will no longer bind to protein kinases in general.

Now why would you want to make this inhibitor not bind to protein kinases? Well now, if it doesn't bind any protein kinase very well because it's too bulky, it doesn't fit anymore, then you can carve an amino acid out of one of the protein kinases by doing homologous recombination or transgenic mutating that particular-- the nucleic acid encoding that gene. And you will have this ability to manipulate both the chemical and the protein target in cases where-- as we'll get to in the last three lectures-- where we're analyzing systems Biology networks, you want to be able to target a particular protein at a time by having a known ligand protein interaction where you minimize the cross plot by engineering the specific interaction.

You start with a specific interaction for the class and then you engineer it so it hits one of them. So that way you can do a time course, say, just knocking out that particular protein quickly or letting it come back. And this shows the results down at the bottom here. You start out with these two different kinases. CDK2, involved in cell cycling. And chem kinase. Two, both of these would bind to the original black inhibitor. And hence, there would be significant crosstalk.

And here, the interfering dosages in micromolars are shown in the three columns here for the three different compounds, run underneath each of them. And you can see that the lower the number, the better it binds. So when you take these-- if you take these now binding pockets and carve them out, that's what the little wedge cut out for the two lowest ones, CDK2, derivative AS1, and chem kinase 2, AS1 derivative.

Now that you can see, they have a much improved binding constant to even the bulkiest derivatives. And this is mediated by a threonine or phenylalanine at position 38 is changed to a glycine. Glycine is obviously smaller than a threonine or phenylalanine at that position and makes room for the drug just in the same way that changing the tyrosine to a phenylalanine made room for the dideoxy terminator in the earlier example.

So in summary, we have talked about protein three dimensional structure and how we can program proteins, basically. How we can use bits of proteins that we may not be able to predict, a priori-- from scratch-- how we get from a sequence to a ligand. But we can take parts that we know and rearrange them in interesting combinations.

We can build up databases of binding constants to combinations of combinatorial libraries of nucleic acids, of peptides, of small molecules. And we can put these together in novel combinations that allow us to do network analysis and ask what protein does what event. So thank you. Until next time.