PROFESSOR: I'm Ernst Frankel. I'll be teaching next two lectures. I'd like to encourage you to contact me outside of class if you have any questions, if you want to meet. And also, please, during class, ask questions. It's a somewhat impersonal setting with the video cameras and the amphitheater, but hopefully we can overcome that.

This unit is going to focus on moving across scales in computational biology, looking from computational issues that deal with the fundamentals of protein structure at the atomic level to the level of protein-protein interactions between pairs of molecules, protein DNA interactions and small molecules, and then ultimately into protein network. So we've got a lot of ground to cover, but I think we'll be able do it. As you've seen in the syllabus, the first couple of lectures are really a detailed look at protein structure, molecular level analysis, and then we'll move into some of these other levels of higher order, including protein DNA interactions and gene regulatory networks.

I think may of you are probably familiar with this quote, that "nothing in biology makes sense except in the light of evolution." And I'd like to offer a modified version of that, which is little in biology make sense except in light of structure, protein structure, DNA structure. We've, of course, seen this very early on in molecular biology when the structure of DNA was solved, and immediately became clear why it was the basis for heredity. But protein structures have been even more lasting impact time and time again, many, many more events, which have really revolutionized the understanding of particular biological problems.

So one example that was stunning at the time had to do with the most frequently mutated protein in cancer. This is the p53 gene. It's mutated in about half of all cancers, and what was observed early on-- this was in the days before genomic

sequencing when it was actually very expensive and hard to identify mutations in tumors.

So they focused on this particular gene, and they observed that the mutations clustered. So this is the structure of the gene from the n-terminus-- the protein from the n-terminus and the c-terminus, and the bars indicate the frequency of mutations. And you can see that they're all clustered pretty much in the center of this molecule.

Now, why is that? It was enigmatic until the structure was solved here at MIT by Carl Pabo and his post-doc at the time, Nikola Pavletich, and they showed, actually, that these correspond to critical domains. And in a second paper, they actually showed why the mutations occur in those particular locations.

So if you look at the plot on the upper left, here's the protein sequence; above it, the frequency of mutations; below it, the secondary structure elements. And you'll see that mutations occur in regions that don't have any regular secondary structure and can occur frequently in regions with secondary structure or not all in regions with secondary structure. So the mere fact that there's a secondary structure element does not define why there're mutations. But when the three-dimensional structure was solved in the complex with DNA, over here on the right-- this is the protein structure on the left, the DNA structure on the right, and in yellow are some of these highly mutated residues.

It turns out that all of the frequently mutated residues are ones that occur at the protein DNA interface. All right, so in a single picture, we now understand what was an enigma for years and years and years. Why are the mutations so particularly clustered in this protein in non obvious ways? Since that is the interface between the protein and the DNA, these mutations upset the transcriptional regulation through the action of p53.

So if we want to understand protein structure in order to understand protein function, where are we going to get these structures from? So the statistics on how proteins themselves-- I show here. This is from the-- I'll call it the PDB, the Protein

Database. Its full name is the RCSB Protein Database, but it's usually just called the PDB. And here, it shows that, at the time of this slide, around 80,000 structures have been determined by x-ray crystallography.

The next most frequent method was NMR, Nuclear Magnetic Resonance, which identified about 10,000 structures, and all the other techniques produce very, very few structures, hundreds of structures rather than thousands. So how do these techniques work? Well, they don't magically give you a structure. Right? They give you information that you have to use computationally to derive the structure.

Here's a schematic of how structures are solved by x-ray crystallography. One has to actually grow a crystal of the protein or the protein and other molecules that you're interested in studying. These are not giant crystals like quarts. They're even smaller than table salt. They're usually barely visible with the naked eye, and they're very unstable.

They have to be kept in solution or, often, frozen, and you should a very high powered x-ray beam through them. Now, most of the x-rays are-- what are they going to do? They're going to pass right through because x-rays interact very weakly with matter. But a few of the x-rays will be diffracted, and from that weak diffraction pattern, you can actually deduce where the electrons were that scattered the x-rays as they hit the crystal.

And so this is a picture, the lower right, of electron density cloud in light blue with the protein structures snaking through it, and what you can calculate, after a lot of work, from these crystallographic diffraction patterns is the location of the electron density. And then there's a computational challenge to try to figure out the location of the atoms that would have given rise to that electron density that then, when hit with x-rays, would have given rise to the x-ray diffraction pattern. So it's actually an iterative process where one arrives at the initial structure and then calculates, from that structure, where the electrons would be, from the position of electrons where the diffraction pattern would be when the x-rays hit it, and determines how well that predicted diffraction pattern agrees with the actual diffraction pattern, and then

continuously iterates.

And so this is obviously a highly computational problem because you not only have to find positions that are maximally consistent with the observed diffraction pattern, but also positions that are actually consistent with physics. So if we have a piece of a molecule here, we can't just put our atoms anywhere. They need to be positioned with well defined distances for the bonds, the bond angles, and so on. So it's a highly coupled problem that we have to solve, and we'll look at some of the techniques that underlie these approaches, although we'll look specifically at how to solve x-ray crystal structures.

I mentioned the second most common technique is nuclear magnetic resonance, and this is a technology that does not require the crystals, but requires a very high concentration of soluble protein, which presents its own problems. And the information that you get out of a nuclear magnetic resonance structure is not the electron density locations, but it's actually a set of distances that tell you the relative distance between two atoms, usually protons, in the structure, and that's what's represented by these yellow lines here. And once again, we've got a hard computational problem where we need to figure out a structure of the protein that's consistent with all the physical forces and also puts particular protons at particular distances from each other.

So we talk about solving crystal structures, solving NMR structures, because it is the solution to a very, very complicated computational challenge. So these techniques that we're going to look at, while not specifically for the solution of crystal and NMR structures, underlie those technologies. What we're going to focus on is actually perhaps an even more complicated problem, the de novo discovery of protein structures. So if I start off with a sequence, can I actually tell you something important and accurate about the structure?

Now, there's a nice summary in a book called *Structural Bioinformatics* that really deals with a lot of the issues around computational biology is relates to structure, that highlights many of the differences between the kinds of algorithms we've been

looking at up until now in this course and the kinds of approaches that we need to take in our understanding of protein structure. So the first and most fundamental obvious thing is that we're dealing with three-dimensional structures, so we're moving away from the simple linear representations of the data and dealing with more complicated three-dimensional problems. And therefore, we encounter all sorts of new problems.

We no longer a discrete search space. We have a continuous search space, and we'll look at algorithms that try to reduce that continuous search space back down to a discrete one to make it a simpler problem. But perhaps most fundamentally, the difference is that now we have to bring in a lot of physical knowledge to underlie our algorithms. It's not enough to solve this as a complete abstraction from the physics, but we actually have to deal with the physics in the heart of the algorithms. And we'll look at the issues highlighted in red in the rest of this talk.

Another thing that's going to emerge is that it would be nice if there was a simple mapping of protein sequence to structures, and if that were the case, you'd imagine that two proteins that are very different in sequence would have different structures. But in fact, that's not the case. You can have two proteins that have almost no sequence similarity at all but adopt the same three-dimensional structure, so clearly, it's an extremely complicated problem made more complicated by the fact that we don't know all the structures. It's not like we're selecting from a discrete set of known structures to figure out what our new molecule is. We have, in potential, infinite number of confirmations and protein chains we need to deal with.

OK, so I hope that you've had a chance to look at the material that I've posted online for review of protein structure. If you haven't, please do so. It'll be very helpful in understanding the next few lectures, and I'll assume that you're familiar with the basic elements, protein structure, what alpha helices are, what beta sheets are, primary structure, secondary structure, and so on.

And I'll also encourage you to become familiar with amino acids. It's very hard to understand anything in protein structure without having some knowledge of what

the amino acids are. The textbook has a nice figure that summarizes the many overlapping ways to describe the features in amino acids, so please familiarize yourself with that.

So these are resources that we posted online. Also, the Protein Databank, the RCSB, has fantastic resources online for beginning to understand protein structure, so I encourage you to look at their website. In particular, in their website, they have tools that you can download to visualize protein structures, and that's going to be a critical component of understanding these algorithms, to actually understand what these structures look like.

I've highlighted, too, that I find particularly easy to use PyMOL and Swiss PDB Viewer. You can not only look at structures with these techniques, you can actually modify them. You can do homology modeling.

So before we get into algorithms for understanding protein structure, we need to understand how protein structures are represented. I've already mentioned that there are these repeating units that I'd like you already know about-- alpha helices, beta sheets. We won't go into those in any detail. But the two more quantitative ways of describing protein structure have to do with a three-dimensional coordinates, the XYZ coordinates of every atom, and internal coordinates, and we'll go through those a little bit of detail.

So again, this PDB website has a lot of great resources for understanding what these coordinates look like. They have a good description of what's called a PDB file, and those PDB files look like this at the outset. They have what is now called metadata, but at the time was just information about how the protein structure was solved. So it'll tell you what organism the protein comes from, where it was actually synthesized if it wasn't purified from that organism, but if it was made recombinantly, details like that, details about how the crystal structure was determined. The sequence-- most of this won't concern us, but what will concern us is this bottom section shown here in more detail.

So let's just look at what each of these lines represents. The lines that contain

information about the atomic coordinates all begin with the word ATOM, and then there's a index number that just is referenced for each line of the file, tells you what kind of atom it is, what chain in the protein it is, and the residue number. So here, it's starting with residue 100. The sequence here can be arbitrary and may not relate to the sequence of the protein as it appears in SWISS-PROT or Gen Bank.

And then the next three columns are the ones that are most important to us, so these are the XYZ coordinates of the atom. So to identify the position of any molecule in three-dimensional space, obviously you need three coordinates, and so those are what those three coordinates are. And they're followed by these two other numbers, which actually are very interesting numbers because they tell us something about how certain we are that the molecule is really-- the atom is really at that position in the crystal structure. So the first of these is the occupancy.

In a crystal structure, we're actually getting the information about thousands and thousands of molecules that are in the repeating units of the crystal, and it's possible that there could be some variation in the structure between one unit of the crystal and the next. So you could have a side chain that, in one crystal, is over here and in the next crystal-- a repeating unit of the crystals over there. If there are discrete confirmations, then you imagine that the signal will be reduced, and you'll actually get some superposition of all the possible confirmations.

So number one here means that there seems to be one predominate confirmation. But if there is more than one, and their discrete-- if they're continuous, it'll just look like noise. It'll be hard to determine the coordinates. But if they're discrete positions, then you might find, for example, an occupancy of 0.5 and then another line with the other position with an occupancy of 0.5. So that's when there's discrete locations where these atoms are located.

The B factor's called the thermal factor, and it tells you how much thermal motion there was in the crystal at that position. Now, what does that mean? If we think about a crystal structure, there'll be some parts of it that are rock solid. In the center, it's highly constrained. The dense core of the protein, not too much is going

to be changing.

But on the surface of the protein, there can be residues that are highly flexible. And so as those are being knocked around in the crystal, they are scattering the x-rays in slightly different ways. But they're not in discrete confirmations, so we're not going to see multiple independent positions. We'll just see some average positions.

And that kind of noise can be accounted for with these B factors, where high numbers represent highly mobile parts of the structure, and low numbers represent very stable ones. A very low number here would be, say, a 20. These numbers of 80-- typically, things like that occur at the ends of molecules where there is a lot of structural flexibility.

So we have this one way of describing the structure of a protein where we specify the XYZ coordinates of every one of these atoms, and we'd have these other two parameters to represent thermal motion and static disorder. Now, are those coordinates uniquely defined? If I have this structure, is there exactly one way to write down the XYZ coordinates?

Hands? How many people say yes? How many people say no? Why not?

AUDIENCE: You can rotate it.

PROFESSOR: You can rotate it. You set the origin. Right? So there's no unique way of defining it, and that'll come up again later.

OK, now, this is a very precise way of describing the three-dimensional coordinates in protein, but it's not a very concise way of representing it. Now, why is that? Well, as the static model represents, there are certain parts of protein structures that are really not going to change very much. The lengths of the bonds change very little in protein structures. The angles, the tetrahedrally coordinated carbon, doesn't suddenly become flat, planar.

These things happen very-- there may be very small deformations. So if I had to specify the XYZ coordinates of this carbon, I really don't have too many degrees of

freedom for where the other carbon can be. It has to lie in a sphere at a certain distance. So instead of representing XYZ coordinates of every atom, I can use internal coordinates.

So here in this slide, we have amino acids-- the amino nitrogen, the carbonyl carbon. So this is a single amino acid. Here's the peptide bond that goes to the next one. And as this diagram indicates, the bond between the carbonyl carbon of one amino acid and the amide nitrogen of the next one is planar, so that angle isn't even rotating. So that's one degree of freedom that we've completely removed.

The angles that rotate in the backbone or called phi and psi; phi over here, and psi over here. So those are two degrees of freedom that determine how this amino acid is-- the confirmation of this amino acid. So instead of specifying all the coordinates, I can specify the backbone simply by giving two numbers to every amino acid, the phi and psi angles, with the assumption that the omega angle, this peptide backbone, remains constant. And similarly for the side chains, and we'll go into this in more detail later, we can then give the coordinates, the rotation, of rotatable bonds in the side chain and not specify every atom as we go out.

OK, so we've got these two different ways of representing protein structure, and we'll see that they're both used. Any questions on this? Great. OK, so if we're looking at protein structures, one question we want to ask is how do we compare two protein structures to each other?

So I already mentioned that proteins can have similar structure, whether or not they are highly similar in sequence. So if I have two proteins that are highly homologous, that do have a high level of sequence similarity-- for example, these two orthologs, this one from cow and this one from rat-- you can see, at a distance, they both have very similar structures. They also have 74% sequence similarity, so that's not surprising. But you can get proteins that have very low sequence similarity. They're still evolutionary related, like these orthologs, two different species that have the same protein, or paralogs, a single species that have two similar copies, but non identical copies, in the same protein that maintain the same structure when they

only have about 20% to 30% sequence similarity.

And you can get even more distant relationships. So here are two proteins, both in human, evolutionarily related, but only 4% sequence identity. And yet at a distance, they look almost identical. And those are evolutionary related proteins, but we can also have things that are called analogs, which have no evolutionary relationship, no obvious sequence similarity, and yet adopt almost identical protein structures. So this adds to the complexity of the biological problems that we're going to try to solve.

All right, so how do I quantitatively compare two protein structures? So the common measurement is something called RMSD, Root Mean Square Deviation, and here, I have a set of structures that were solved by NMR. And you can see that there's a core of the structure that's well determined and then there are pieces of the structure that are poorly determined. There weren't enough constraints to define them.

And these proteins have all been aligned, so the XYZ coordinates have been rotated and translated to give maximal agreement. And what's the agreement measure? It's this Root Mean Square Deviation.

So I need to define pairs of atoms in my two structures. If it's, in this case, the same structure, that's really easy. Every atom has a match in this structure that was solved with the same molecule.

But if we're dealing with two homologous proteins, then that becomes a little bit more tricky. We need to define which amino acids are going to match up. We can also define whether we care about changes in the side chains, or whether we only care about changes in the backbone, whether we're going to worry about whether the protons in the right places or not. And you'll see that these alignments can be done with either only heavy chain, heavy atoms, meaning excluding the hydrogens, or only main chain atoms, meaning excluding the side chains completely.

But once we've defined the pairs of corresponding atoms, then we're going to take the difference in the distance squared, sum of the squares of the distances between

the corresponding atoms and their x-coordinate, their y-coordinate, and they're z-coordinate. Take the square root of that sum, and that's going to give us the Root Mean Square Deviation. And of course, we have to minimize that Root Mean Square Deviation with these rigid body rotations to account for the fact that I could have my PDB file with the origin of this atom. Or I could have my PDB file with the origin of that atom, and so on.

OK. Any questions so far? Yes.

AUDIENCE: Do we consider every single atom in the molecule?

PROFESSOR: So we have a choice. The question was do we consider every single atom in the molecule? We don't have to do, and it depends, really, on the problem that we're trying to solve. So if we're looking for whether two proteins have the same fold, we might not care about the side chains. We might restrict ourselves to main chain atoms.

But if we're trying to decide whether two crystal structures are in good agreement with each other, or say, as we'll see a few minutes, we're going to try to predict the structure protein, and we have the experimentally determined structure of the same protein, and we want to decide whether those two agree, in that case, we might actually want to make sure that every single atom is in the right position. So it'll depend on the question that we're trying to answer. Good question.

Any other questions? OK. All right, so so far, I've shown a lot of static pictures of molecules. I do want to stress that molecules actually move around a lot, so I'll just show a little movie here.

[VIDEO PLAYBACK]

[END VIDEO PLAYBACK]

PROFESSOR: OK, so that was, in part, an excuse to play a little New Age music in class, but more fundamentally, it was to remind you that, despite the fact that we're going to show you a lot of static pictures of proteins, they're actually extremely dynamic. And they

have well defined structures, but they may have more than one well defined structure, especially those molecules that are doing work. They're actually moving things along. They have multiple structures. And so when we consider the protein structure, it's an approximation, and we're always going to mean the protein structures, not singular one.

OK, so what determines the protein structure? Well, I've told you it's physics. Fundamentally, it's a physical problem, so the optimal protein structure has to be an energetic minimum. There has to be no net force acting on the protein.

The force is negative derivative of the potential energy, so that derivative has to be 0. So we have to have a minimum of protein structure. Now, that doesn't mean that there's exactly one minimum.

Those proteins that had multiple confirmations in that movie obviously had multiple minima that they could adopt depending on other circumstances, but there has to be at least a local minimum. So if we knew this U, this potential energy function, and we could take the derivative of it, we could identify the protein structure or the protein structures by simply identifying the minima in that potential energy function. Now, would that life were so simple, right?

But we will see that there are ways of parameterizing the U and using it to optimize the structure so it finds this, at least local, minimum. And we're going to look primarily at two different ways of describing the potential energy function. One of them, we're going to look at the problem like a physicist one, and the other way, we're going to look at it as a statistician would.

So the physicist wants to describe, as you might imagine, the physical forces that underlie the protein structure, and so as much as possible, we're going to try to write down equations that represent those forces. Now, we're not always going to be able to do that because a lot of forces involved are quantum mechanical. The mere fact the two solid objects don't pass through each other is because of exclusion principles that deal with quantum mechanics.

We're not going to write down quantum mechanical equations for every atom in our protein structure, but we will write down equations that approximate those. And wherever possible, we're going to try to tie the terms in our equations into something identifiable in physics, and a very good example of this approach is the CHARMM program. And these approaches actually were the ones that won the Nobel Prize in chemistry this past year.

At the other end of the spectrum are the statistical approaches. Here, we don't really care what the underlying physical properties are. We want equations that capture what we see in nature.

Now, often, these two approaches will align very well. There'll be some approximations that the physicist makes to capture a fundamental physical force. That's simply the best way to describe what you see nature, and so those two terms may look indistinguishable in the CHARMM version or my favorite statistical approach, which is Rosetta.

So we'll see that some terms in these functions agree between CHARMM and Rosetta. Well, there'll be places where they fundamentally disagree on how to describe the molecular potential energy function because one is trying to describe the physical forces and the other one is trying to describe the statistical ones. Do we have any native speakers of German in the audience?

AUDIENCE: I'm a speaker.

PROFESSOR: You want to read the joke for us?

AUDIENCE: Yeah. Institute for Quantum Physics, and it says "You can find yourself here or here."

PROFESSOR: OK.

AUDIENCE: [LAUGHTER]

PROFESSOR: All right, so for the video, it's the Institute for Quantum Mechanics. And you go to a map at MIT, and it'll say, you find, "You are here." Right? But in the Institute for

Quantum Mechanics, it says "You're either here or here."

So that's the physicist approach. We really do have to think about those quantum mechanical features, whereas on the right-hand side is the statisticians approach. It says "Data don't make any sense. We'll have to resort to statistics." OK? So the statistician can get pretty far without understanding the underlying physical forces.

All right, so let's look at this physicist approach first, so we're going to break down the potential energy function into bonded terms and non-bonded terms. So the bonded terms, as they sound, are going to be atoms that are close to each other in the bonded structures, so certainly these two atoms, because they're connected by a single bond, are going to be bonded terms. But we'll see groups of three or four atoms near each other will also be bonded terms. And the non-bonded terms will be when I have another molecule that comes close, but isn't directly connected. What are the physical forces between these two ?

So these bonded terms then first break down into a lot of sub terms. I'll show you the functional forms here. We'll just look at a few of them in detail and then give you a sense of what the other ones are.

So this first one is the bonded term that describes, actually, the distance between two bonded atoms. Now, again, this is fundamentally quantum mechanical property, but it would be too computationally expensive to describe the quantum mechanics and not really necessary because you can do pretty well by just describing this as a stiff spring. So that's what this quadratic form of the equation represents.

So we simply define b naught here as the equilibrium position between these two atoms, particular types. There would be two tetrahedral coordinated carbons, and that would be determined by looking at a lot of very, very high resolution structures in small molecule crystals so we know what the typical distance for this bond is. We get that as a parameter. There would be a big file in the CHARMM program that lists all those parameters for every one of these bonded terms, and then if there's a small deviation from that, because the molecules stretched a bit in your refinement process, there would be a penalty to pull it back in just like a spring pulls it back in.

Now, it turns out that when you go this route, you have to actually come up with a lot of equations to maintain the geometry because, again, we're going to have to not only worry about these distance bonds, but we need to worry about angles. So we've got the angle between this bond and this bond. What keeps that in place?

So we need to add another term that's a second term here to make the angle between these fixed, and then we have to deal with what are called dihedral angles to make sure that these four atoms lie in the allowed geometry. And so each one of these terms accounts for something like that. This last term over here makes sure that the phi and psi angles are consistent with what we see in quantum mechanics as corrected for any deviations that we see in these small molecules so a lot of terms with a lot of parameters they're trying to capture the best description of what we observe in each one is motivated by the fact that there is some quantum mechanical principle underlying it. So-- yes?

AUDIENCE: Why is the [INAUDIBLE]?

PROFESSOR: I actually don't know the answer to that. But there's a reference there that I'm sure will give you the answer. OK, now what about these non-bonded terms? So non-bonded terms of the set are molecules that are distant from each other in the structure of the protein, but close to each other in three-dimensional space. And there are two fundamental forces here.

The first one is called the Leonard Jones potential, and the second one of the electrostatic one. And the Leonard Jones potential itself has these two terms. One is an R6 term, a negative r to the 6th dependency. The other one is positive nr to the 12th.

The negative r to the 6th is an attractive potential. That's why it's negative, and it's because of small induced dipoles that occur in the electron clouds of each of these atoms that pull the molecules together. And the 1 over r to the 6th dependency has to do with the physics of two dipoles interacting.

The r over 12 term is an approximation to a quantum mechanical force. So the reason the two molecules don't pass through each other, as we said already, is because quantum mechanical forces. That would be very expensive to compute, so we come up with a term that's easy to compute. And of course, an r 12 term is simply the square of an r to the 6th term, so if you already computed 1 over r to the 6th between two atoms, you just square that, and you get 1 over r to the 12th. So it's very computationally efficient, and you adjust the parameters, these r mins, so that it works out so that these things agree reasonably well with the crystal structures. And these are crystal structures of small molecules that we know in great detail.

And then the electrostatics is what you might expect for electrostatics. It's got a potential that varies as 1 over the distance, and as the product of those charges, these can be full charges or they can be partial charges. And there's a term here, this epsilon, which is the dielectric constant, and that represents the fact that, in vacuum, there'd be much greater force pulling two oppositely charged molecules together than in water because the water's going to shield. And so these electrostatic terms, this dihedral dielectric potential term, can vary from one, which is vacuum, to, say, 80 for water. And setting that is a bit of an art.

OK, so what do these potentials look like? Those are shown here. This is the, in dark lines, the sum of the van der Waals potential. It consists of that attractive term, which has the r over 6 dependency, and the repulsive term with the r over 12. And why does it go up so high at short distances?

AUDIENCE: [INAUDIBLE].

PROFESSOR: Right, because you can't have molecules that overlap. You'll see that there's a minimum, so there's an optimal distance barring any other forces between two atoms. So that's roughly what these hard sphere distances represent in the scale models. And then the electrostatic potential also, obviously, has attractive term, but it's going to blow up as you get to small values, increasingly favorable.

And so the net sum of those two is shown here, the combination of van der Waals

and electrostatics. It, again, has a strong minimum but becomes highly positive as you get to close distances. OK, any questions on these forces? Yes?

AUDIENCE: Do the van der Waals equal the Leonard Jones potential? Or is that something else?

PROFESSOR: Yeah, typically, those two terms are used interchangeably. Yeah. Other questions? OK.

All right, so that's how the physicist would describe the potential energy function. Rosetta, as I told you, is an example of the statistical approach. It rejects all this sharp definition of trying to compute exactly the right distance between two atoms by having a stiff spring between them and says let's just fix a lot of these angles.

So we're going to fix the distance between two atoms. There's no point in having them vary by tiny, tiny fractions in the bond length. We're going to fix a tetrahedral coordination of our tetrahedral carbons. We're not going to let them deform because that never would happen in reality, and so we're going to focus our search over the space entirely over the rotatable bonds.

So remember, how many rotatable bonds did we have in the backbone? We had two, right? We had the phi and the psi angles, and then the side chains then will have rotatable bonds over the side chains.

So in this example, this is a cysteine. Here's the backbone. Here's the sulfur. And we have exactly one rotatable bond of interest because we don't really care where the hydrogen is located.

So we've got this chi 1 angle. If there were more atoms out here, this would be called chi 2 and chi 3. And these can rotate, but they don't rotate freely. We don't observe, in crystal structures, every possible rotation of these angles, and that's what this plot on the left represents.

For this side chain, there's a chi 1, a chi 2, and a chi 3, and the dark regions represent the observed confirmations over many, many crystal structures. And you

can see it's highly non uniform. Now why is that?

I see people with their hands trying to figure it out in the back. So why is that? Figure that's what you guys are doing. If not, it's very interesting sign language.

So if we look down one of these tetrahedral carbon-carbon bonds, we have apparently a free rotation. But in fact, some these confirmations, we're going to have a lot of steric clashes between the atoms on one carbon and the atoms on the other, and so this is not a favorable confirmation. The favorable confirmation is offset, and that propagates throughout all the chains in the protein.

So there'll be certain angles that are highly preferred, and other ones that are not. These highly preferred angles are called rotamers, and so we'll use the term a lot. It stands for rotational isomers.

And so now, we've turned our continuous problem of figuring out what the optimal angle is for this chi 1 rotation into a discrete problem where maybe there are only two or three possible options for that rotation. And so now, we can decide is this better than this one or this one? Questions on rotamers or any of this? Excellent.

OK, so how do we determine-- we've decided then we're going to describe the protein entirely by these internal coordinates-- the phi, the psi, the backbone, the chi angles of the side chain. We still need a potential energy function, right? That hasn't told us how to find the optimal settings, and we're going to try to avoid the approach of CHARMM, where we actually look at quantum mechanics to decide what all the terms are. So how do they actually go about doing this?

Well, they take a number of high resolution crystal structures, and they characterize certain properties in those crystal structures. For example, they might characterize how often a certain aliphatic carbon-- how often aliphatic carbons are near amide nitrogens, and they might measure the distance-- they do measure the distance between these amide nitrogens and aliphatic carbons across all the crystal structures and determine how often those distances occur. And you can actually turn those observations, then, into a potential energy function by simply using

Boltzmann's equation. So we can figure out how frequently we get certain distances on the x-axis is distance, on the y-axis is frequency, number of entries in the crystal structure, and then by Boltzmann's Law, we can compute the density of states over some reference, which is actually very hard to define. And you can look at some of the references referred to in the slides to figure out how currently that's defined, but we have to find some arbitrary reference state to figure out the probability of being any one of these states is going to be a function, a logarithmic function, of the frequency of those states.

All right, so we've got an energy term that's determined solely by the observations of distances, that doesn't say I know that this one's charge and this one isn't. It just says here's an oxygen attached to a carbon with double bonds. Here's a carbon that's not. How often are they at any particular distance? And we go through lots and lots of other properties, and we'll go into detail now to what those other terms are to look through high resolution crystal structures, see what certain properties are, turn those into potential energy functions that we can then use to identify the optimum rotations for the side chain and the backbone.

Oh, and I should also point out that when we do this, we'll have different terms for different things. We'll have a term for distances between different kinds of atoms. We'll have terms for some of these other pieces of potential energy that we'll describe in subsequent slides, and we're going to need to decide how to weight all of those, all those independent terms, to get them to give us reasonable protein structures when we're done. And that, once again, is a curve fitting exercise, finding the numbers that best fit the data without any guiding physical principle underneath it.

So you'll be using PyRosetta. And in PyRosetta, you'll see the terms on the board for the potential energy functions, the different features of the potential energy function, and I'll step you through a few of these just so you know what you're using. There'll also be files in PyRosetta installation that will give you the relative weights for each of these terms.

OK, so these first are the van der Waals, and here, the shape of the curve looks just like we saw before. It has to, in some sense because they're trying to solve the same physical problem, but the motivation is very different. There's no attempt to decide that it should be a 1 over r to the 6th because of dipole-dipole interactions. And simply, how do I find the function that accurately represents what I see in the database? So again, computed, this is the fa attractive and the fa repulsive, and those are determined based on the statistics of what's observed in the crystal structures.

This one, the hbond, breaks down into backbone and side chain, long range and short range. And the goal of the hbonds-- so hydrogen bonds are one of the principal determinants of protein structure, and you'll see that in the reading materials that are posted online. And one of the critical things about a hydrogen bond is that it needs to be nearly planar. So the line between-- the angle between this atom, which has the hydrogen attached, and this one, which is the free electron pair, has to be as close to linear as possible. And the more it deviates from linear, the weaker the hydrogen bond will be.

And so this hydrogen bonding potential has terms that describe the distance between the atoms that are donating and accepting the hydrogen as well as the angle between them, and it's been parameterized to represent, separately, things that are far from each other, close to each other, things that are side chain, or main chain. And here's where it's really the statistician against the physicist. Why divide up side chain and main chain? There's no physical principle that drives you to do that. It's simply because that's what gives the best fit to the data, so the statistician is not afraid to add terms that make their models better fit reality, even if they don't represent any fundamental physical principle.

And we'll see it gets even more dramatic with some these other terms. So this is the Ramachandran plot, which you'll also see in your reading. It represents the observed frequencies of phi and the psi angles. And as you know that there are only a couple positions on this phi and psi plot that are frequently observed, representing the different regular secondary structures primarily, alpha helix and beta sheet is

indicated.

And rather than trying to capture the fact that protein should form alpha helices by having really good forces all around, they simply prefer angles that are observed in the Ramachandran plot. So we're going to give a potential energy function that's going to penalize you if your phi and psi ends up over here, and reward you if your phi and psi ends up in one of these positions. So from the physicist, this is cheating, and for the statistician, it makes perfect sense. Shouldn't laugh at that.

OK, and this same will be true for the row numbers. So we said that, for the side chains, there are certain angles that we prefer over others because that's what we observe in the database. Again, we're not going to try to get them by making sure that there's repulsion between these two atoms when they're eclipsed. We're going to get there simply by saying the potential energy is lower when you're in one of these staggered confirmations than you're one of the eclipse confirmations.

OK, now, the place where the difference between the statistician and the physicist is most dramatic comes when we look at the salvation terms. So a lot of what goes on in protein structure-- determines protein structure, I should say, is the interaction of the protein with water. It's bathed in a bath of 55 molar water molecules, highly polar. They normally are hydrogen bonding with each other. When the protein sits in there, the protein has to start hydrogen bonding with them.

And where do we find hydrophobic residues in a protein structure, with your hands? Outside or inside? Inside, right? So the hydrophobic residue's all going to be buried inside. Why is that?

Well, it's actually really, really hard to describe in terms of fundamental physical principles. In fact, it's really hard to describe the structure of water by fundamental physical principles. Simulations that try to get water to freeze were only successful a few years ago. So we've tried to simulate water using basic physical principles. It's very hard to get it to form ice when you lower the temperature, so it's going to be even harder, then, to represent how a complicated protein structure immersed in the water actually interacts with those water molecules.

So you've got all these water molecules interacting with polar residues or non-polar residues. The physicist really struggles to represent those. And just to show you why that is, let me show you, again, a little movie. Unfortunately, no new age music with this one. I apologize.

So what's shown here is a sphere immersed in a bunch of water molecules. The red is the oxygen. The little white parts are the hydrogens. You can see them wiggling around.

And what's the fundamental feature that you observe? All right, they're forming almost a cage around this hydrophobic molecule. Why is that? Yeah?

AUDIENCE: It's hard for them to interact with a non-polar residue.

PROFESSOR: Right, so it's hard for them to interact with a non-polar residue. So the water molecules want to minimize their potential energy. They're going to do that by forming hydrogen bonds with something. In bulk solvent, they form it with other water molecules.

Here, they can't form any hydrogen bonds with a sphere, so they have to dance to this complicated dance to try to form hydrogen bonds with each other with this thing stuck in middle of them. And this is, at its heart, the fundamental driving force between the hydrophobic effect, that which causes the hydrophobic residues to be buried inside of the protein. Very, very hard, as I said, to simulate using fundamental physical forces.

So what does the statistician do? The statistician has a mixture of experimental observation and statistics at their benefit, so we can measure how hydrophobic any molecule is. We can take carbons and drop them to non-polar solvents, into polar solvents, and determine what fraction of time a molecule will spend in a polar environment versus a non-polar environment, and from that, get a free energy for the transfer of any atom from a hydrophobic environment to a hydrophilic environment. That can give us is delta G Ref, shown over here.

OK, now, in a protein, that molecule is not fully solvent exposed even when it's on the surface, because water molecules trying to come at it from this direction can't get to it, from this direction can't get to it. So the transfer energy for this carbon to go from fully solvent exposed to buried is different from the isolated carbon. And so the statistician says, OK, I'll come up with a function to describe that. I will describe what else is near this atom in the rest of the protein structure.

That's what the term on the right does. It's a sum over all other neighboring atoms and describes the volume of the neighboring group. Is the thing next to it really big or really small? Usually not described, necessarily, at the level of atoms. It might be side chains depending on which program is doing it.

But I have some measure of the volume of the neighbors. If that volume is really large, then this thing is already in a hydrophobic environment even when it's taking water because it's surrounded by bulky things. If the neighbors are small, then it's a more hydrophilic environment when it's taking in water, and that's going to modulate this free energy. Is this function clear?

OK, so by combining this observation from small molecule transfer experiments and these observations based on the structure of the protein, we can get an approximation for the hydrophobic effect. How expensive is it to have this piece of the protein in solvent versus in the hydrophobic core? And again, we never had to do any quantum mechanical calculations.

We never had to actually explicitly compute the interaction of this molecule with solvent. We don't need any water in the structure. It's simply the geometry of the protein that's going to give us a good approximation to the energy function.

All right, so you can look through all the details of these online in the Rosetta documentation that we provided to get a better sense of what all these functions are, but you can see there are a lot of terms. It's increasingly incremental. You find something wrong with your models. You add a term to try to account for that. Again, not driven necessarily by the physical forces.

OK, so what have we seen so far? We've seen the motivation for this unit, to begin with protein structures, that the protein structure really helps us understand the biological molecules that we're looking at. These structures are going to influence our understanding of all biology, so we need to be good at predicting these protein structures or solving them when we have experimental data. The computational methods that we're going to use-- we're going to focus on solving protein structures de novo, predicting them, but those same techniques are going to underlie the methods that are used to solve x-ray crystallography in an MR.

And fundamentally then, we have these two approaches to describing the potential energy. That's the statistician and the physicist's approach. And remember, the key simplifications of the statistician are that we used a fixed geometry.

We're not trying to figure out the XYZ coordinates of every atom. We're simply trying to figure out the bond angles. We're going to use rotamers, so we're going to turn our continuous choices often into discrete ones. And we're going to derive statistical potentials to present the potential energy, which may or may not have a clear physical basis.

All right, so let's start with a little thought experiment as we try to get into some of these prediction algorithms. So I have a sequence. It's about, I don't know, 100 amino acids long, and here are two protein structures. One is predominantly alpha helical. One is predominantly beta sheet.

How could I tell-- this is not a rhetorical question. I want you to think for second. How could I tell whether the sequence prefers the structure on the top or the structure on the bottom? So we have, actually, a lot of the tools in place. Yes, in the back.

AUDIENCE: Can you, based on previously known sequences, know which sequence is predominant in which [INAUDIBLE]?

PROFESSOR: OK, so the answer was we could look at previously known sequences. We can look for homology, and that's actually going to be a very powerful tool. So if there is a

homologue in the database that is closely related to this protein, and it has a known structure, then problem solved. What if there isn't? What's my next step? Yes?

AUDIENCE: What if you start with a description of the secondary structure, say the helices and the sheet, and you counted how often a particular amino acid showed up in each of those structures? Could you then compute maybe a likelihood across a stretch of amino acids?

PROFESSOR: Great. So that answer was what if I looked at these alpha helices and beta sheets and computed how often certain amino acids occur in alpha helices versus beta sheets, and then I looked in my protein structure and checked whether I have the right amino acids that are more favorable than alpha helices or beta sheets. And we'll see that's an approach that's been used successfully. That's secondary structure prediction. OK, other ideas. Yep?

AUDIENCE: So if you have the position of the 3D structure, you can feed your sequence through the structure and then put it through your energy function, see which one is the lower [INAUDIBLE].

PROFESSOR: Excellent. So another thing I can do is, if I have these two structures, I have their precise three-dimensional structures, I could try to put my sequence onto that structure, actually put the right side chains for my sequence into that backbone confirmation. And then what would I do? I would actually measure the potential energy of the protein in top structure and the potential energy of the protein in the bottom structure.

If the potential energy is higher, is that the favorable structure or the unfavorable structure? Favorable? Unfavorable? Right, it's the unfavorable. So I want the lower free energy structure.

OK, so let's think about-- that's correct, and that's where we're headed. But what are going to be some of the complexities of that approach? So first of all, what about these side chains? I have to now take a backbone structure that had some other amino acid sequence on it, and I have to put these new side chains on. Right?

If I put those on in the wrong way-- let's say, this is the true one-- let's say one of these is the true structure. Let's begin with a simplification. All right, so let's say your fiendish labmate has actually solved the structure of your protein, but refuses tell you what the answer is.

AUDIENCE: [LAUGHTER]

PROFESSOR: And she actually has solved two structures, neither one of which she's going to give you the sequence to. But she's giving you the coordinates for both of them. They're the same length.

And so she asks you, ha, you took 791. You can figure this out. Tell me whether that your sequence is actually in this structure or that structure. She says one of them is exactly right. You just don't know which one.

OK, so she gives you the backbone coordinates, so you go. You put your amino acid sequence, say, with Swiss [? PDB. ?] You add to the backbone all the right side chains. But now, you have to make a bunch of decisions for these side chain confirmations. If you make the wrong decision, what happens?

Well, you stick this atom close to where some other atom is. Now, you've got an optimization problem, right? You believe that one of these backbone coordinates is correct, but you've got a very highly coupled optimization problem.

You need to figure out the right rotations for every single side chain on this protein, and you can't do it one by one. You can't take a greedy approach because if I put this side chain here, and I put this side chain here, they collide, but if this was wrong and supposed to be over there, then maybe this is the right conformation. So I have a coupled problem, so it turns out to be computationally expensive thing to compute. So we're going to look at what to do if we know backbone confirmation, but we don't know the side chain confirmation. We can try to solve that optimization problem, and you'll actually do that in your problem set.

Now, what if the backbone confirmation isn't exactly correct? So let's say you do what was first suggested, and you search the sequence database. You take this

sequence, and you find that it actually has two homologs, two things with similar sequence similarity. There are two proteins with 20% sequence identity that have completely different structures.

This one has 20% sequence identity, and this one has 20% sequence identity. So you have no way of deciding which one's which, right? And neither one is going to be the right protein structure.

So you know that by putting the side chains onto these protein structures, you do have to solve those problems with side chain optimization, but what, obviously, is the other thing that you're going to need to have to solve? All right, you're going to need to solve the backbone optimization problem, and this becomes even more coupled because when I move this backbone, then the side chains move with it. So now, I've got a very, very complicated optimization problem to deal with. The search space is enormous, and even if I discretize it, it's still very, very large. In fact, there's something famous called the Levinthal Paradox.

Of course, Cy Levinthal, who was once upon a time a professor here and then moved to Columbia-- he did a back of the envelope calculation for extremely simple models of protein structure. If you imagine the proteins were to randomly search over all possible confirmations with very rapid switching between possible confirmations, it would take basically the lifetime of the universe for a protein to ever fold. So proteins don't do random searches over all possible confirmations, and they can check out confirmations incredibly rapidly. So we certainly can't do that, so we'll look at the optimization techniques.

All right, so we discussed how to use energy optimization functions to try to decide which one's correct, and that even if the structure is the correct one, we have the side chain optimization problem. If the structure's the incorrect one, we've got two problems. We've got the backbone confirmation and the side chain.

This is frequently called fold recognition or threading. This choice of, you've got a protein structure. You want to decide if your sequence matches this one or that one.

There are a couple of other problems that we're going to look at. So this was already raised by one of the students, the idea that we try to predict the secondary structure of this protein, so we'll look at secondary structure prediction algorithms. This was a very early area of computational effort in structural biology, and we'll see that the early methods are remarkably good.

We can look for domain structures, and this is really a sequence problem. So we can look through our sequences, and rather than looking for sequence identity or similarity with known structures, we can see whether there are certain patterns, like the hidden Markov models that you looked at in a previous lecture, that can allow us to recognize the domain structure of a protein even without an identical sequence in the database. So we won't go over that kind of analysis anymore, and then we'll spend a good amount of time looking at ways of solving novel structures. So if you don't have a fiendish friend who solved your structure for you, and there is no homologue in the database, all is not lost. You actually can now predict novel structures of proteins simply from the sequence.

All right, so a little history as to the prediction of protein structure. It really starts with Linus Pauling, who went on to win the Nobel Prize for this work. And this is in the era-- this paper was published in 1951. This was what computers looked like in 1951, and that thing probably has a lot less computing power than your iPhone or your Android.

So Linus Pauling did not solve the structure of the alpha helix, predict that alpha helices existed, using computers. He actually did it entirely with paper models. And in fact, he solved this-- he got the key insights for the alpha helix when he was lying sick in bed. That's a very productive sick leave, you might imagine.

He was using paper models, but it wasn't all done while lying in bed. So he and others, the field as a whole, have spend a lot of time observing small molecule distances, so they have some idea what to expect in protein structures. They didn't know the three-dimensional structure, but they knew a lot of the parameters about how far apart things were. And they also knew that hydrogen bonds were going to

be extremely favorable in protein structures.

And so he looked for a repeating structure that would maximize the number of hydrogen bonds that occur within the protein backbone chain. And he knew, also, the backbone-- that the amide bonds would be planar and so on. So there were a lot of principle that underlay this, but it was really a tour de force of just thinking rather than computing.

Another really important contribution early on was made by Ramachandran, was at Madras University, and his insight had to do with the fact that not all backbone confirmations were equally favorable. So remember, we have these two rotatable bonds in the backbone. We have the phi angle and the psi angle. And this plot shows that there'll be certain confirmations of phi and psi angles that are observed within these dashed lines, and then the other confirmations, which are almost never observed.

Now, how did he figure that out? Once again, it wasn't with computation. It was simply with paper models and figuring out what the distances would be, and then carefully reasoning over those possible structures. So you can get very far in this field, initially, back then, by simple hard thought.

OK, so with these two observations, we knew that there were going to be certain kinds of regular secondary structure and that not all backbone confirmations were equally favorable. OK, but now, we want to advance actually predicting structures of particular proteins, not just saying that proteins in general will contain alpha helices. So how do we go about doing that?

So the first advances here, we're trying to predict the structure of alpha helices, and this paper in the 1960s introduced the concept of a helical wheel. Now, the idea here, if you'll imagine that this eraser is an alpha helix, I'm going to look down the backbone of the alpha helix. And I'll see that the side chains emerge at regular positions. There's going to be 100 degree rotation between each sequential residue in the backbone as it goes around helix. It's going to be displaced and rotated by 100 degrees, and I could plot, on a piece of paper, the helical projection, which is

shown here.

So here's the first amino acid. 100 degrees later, the second. 100 degrees later, the third. And I can ask whether the residues on that backbone have a sequence that puts all the hydrophobics and hydrophilics on the same side, as in this case, or on different sides.

Now, what difference does it make? Well, if I have an alpha helix that's lying on the surface of a protein, this could have one side that's solvent exposed and one side that's protected. So we would expect that some of these alpha helices lying on the service would be amphipathic. Half of them would be hydrophobic, hydrophobic, and half of them would be hydrophilic. And purely, as someone suggested from the pattern of the amino acids, and here the hydrophobicity of the pattern of the amino acids, we could make reasonable predictions of whether this protein forms a particular kind of alpha helix, an amphipathic alpha helix.

Now, is that going to help us for all alpha helices? Obviously not, because I can have alpha helices that are totally solvent exposed, and I can have alpha helices that are totally protected. So this pattern will occur in some alpha helices, but not all.

So another idea that was raised here and was used early on with great success was to actually figure out whether certain amino acids have a particular alpha helical propensity. Do they occur more frequently in alpha helices? At the time, it was also thought maybe you could find propensities for beta sheets and other structures.

So compute the statistics over for every amino acid, shown as a row here. How often is it observed in the database? How often does it occur in alpha helix? And how often does it occur in beta sheet or in a coil? And from these, then, we would compute probabilities and compute using, perhaps, Bayesian statistics to compute the poster expectation for having a certain sequence in alpha helix.

They didn't quite use Bayesian statistics here. They came up with a rather ad hoc approach, and when you read it in hindsight, it seems kind of crazy. But actually, you have to remember when this was being done. This is being done before a big

influence of mathematicians into structural biology. This is 1974, and they used more physical reasoning.

They knew something about how alpha helices formed from chemistry. They knew that, typically, there's nucleation event, where a small piece of helix forms initially, and then that extends. They knew that there were these propensities for certain amino acids to form alpha helices, and other amino acids, which tended to break the helix. And they came up with an ad hoc algorithm that counted how often you had strong helix formers, how often you breakers. You can see all the details in the references.

The amazing thing is, with this very ad hoc thing and a very, very small database of protein structures, you could look at the total number of residues that they're looking at over all the structures, there's 2,473 and residues, not structures. And now, we have many, many more times than that of structures of proteins. Even with that, in 1974, they were able to achieve 60% accuracy in predicting the secondary structure of proteins, so it's really an astounding accomplishment.

And to put that in perspective, there was an evaluation of a whole bunch of secondary structure prediction algorithms done about a decade ago, and things haven't changed that much since then, where between 1974 and 2003, almost 30 years, they went from 60% accuracy to 76% accuracy. OK, well, that's not bad, but it's not a lot for-- you'd expect maybe over 30 years, you could do a lot better. So the simple approach really captured the fundamentals of predicting secondary structure. There's a lot of work that's been done since, and I encourage you to look in the textbook if you're interested, to look at all the newer algorithms that have tried to solve the secondary structure prediction problem. OK.

All right, so secondary structure prediction, then-- you can look in the textbook for the modern methods, but the fundamental ideas were laid down by Chou and Fasman in the 1974 paper. We're already said that looking at the kinds of approaches that we discussed earlier in the course can help you solve domain structures. I would like to focus on, at the end of this lecture and the beginning-- and

the next lecture about how to actually solve novel structures from purely amino acid sequence, and we're going to go back to the idea that there is a potential energy function.

We now have both the CHARMM approach and the Rosetta approach to protein structure, and so there is some protein folding landscape. There's an energy function. If you have different conformations, you'll be at different positions in landscape, and we'd like to figure out how to go from some starting confirmation that may be arbitrary and find our way to the minimum energy structure.

All right, so there are going to be three fundamental things that we'll talk about in the next lecture. We're going to talk about energy minimization, how to use these potential energy functions that we started off with to go from approximate structures to the refined structure. That's the thought problem I gave you.

You have the structure, but you have the wrong side chains. Could you minimize them? And so that's making small changes.

We'll discuss molecular dynamics, which actually tries to simulate all the forces on a protein and to actually carry out a physical simulation of the process. That's the CHARMM approach, and we'll see some interesting variants on that. And then we'll look at simulated annealing, which is an optimization technique that's actually quite broad, but can be applied here, to search over large, large conformational spaces, much further than a protein would actually evolve in a molecular dynamic simulation that's simulating protein function.

You allow the protein, now, to jump between confirmations that have no real potential to transfer between in a normal room temperature in water, but can be done, obviously, easily in the computer. So I'll stop here. Any questions before we close?