NARRATOR: 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 ocw.mit.edu.

DR. GEORGE CHURCH: OK, welcome back. One quick announcement before we get started. The teaching Fellows and I have, in response to a few gentle inquiries have, come up with a plan for making your problem sets and projects a little bit easier, getting you some time. Take a look at the website and talk to your teaching Fellows. But basically, there will be no problems in set six.

It will be combined with problem set five, and so you'll have a full three weeks to work on your project. Tripling your project time. Assuming you haven't done anything so far, which hopefully all of you have done a lot so far. OK, so we were correlating absolute levels of messenger RNA abundance in messenger RNA protein.

And here, this study was subjected to a little bit of not terribly controversial critique. What you're seeing here is at very-- if you have the very low abundance proteins, you might have what looks like some correlation. Then you add a few more low abundance proteins, so this is just some fluctuation. And adding proteins should improve your correlation coefficient.

But if they're of low reliability in either the protein of the messenger RNA scale, or if there isn't correlation between the two for biological reasons, then your correlation coefficient could drop just at random. But as you add-- as you get up to adding all the proteins, you could either be dominated by a few high abundance proteins that fit this perfectly, the messenger RNA protein abundance.

Anyway, some of the critiques of this had to do with the assumptions underlying the Pearson linear correlation coefficient, which in calculating the statistical significance of some of these analyzes in the previous slide, makes an underlying assumption that you have a normal or Gaussian bell curve. And you don't need to do this in order to-- the variety of measures of correlation, which do not require this parametric assumption.

And there are tests for how close to normal a distribution is. Deviations from normality can of all types. For example, you can have slightly flatter or slightly sharper than normal. You can have skewed to the left or to the right, and so forth. Some of these, like skewing, can be corrected by a log transformation, where you simply take the logarithm or some other transformation, and a log being by far the most common, and theoretically justified, and now it becomes normal, and then you can do a statistical test.

As it turns out, this is not a huge effect, but they use it in order to point out that you can when you're testing these, especially the low abundance end of the spectrum. You might want to use a rank test. And a rank test here is illustrated that you take if you have, say, two columns, this is a series of pairs of intensities of messenger RNA and proteins. Say columns X and column Y down in the lower right-hand corner of Slide 37.

And let's say the abundances, the absolute abundances for protein X is 1, 6, 6, and corresponding RNA Y, which would be 8, 2, 3, and 4. Now, you want to ask whether the correlate or not. And what you do is rank them, and so the rank of X is 1, 3, 3. Here is a tiebreaker-- the way you deal with a tie is you give them all the rank of the middle one of the series in the pi.

And you get a rank for Y, and the total number, in this case, is 4, and so when you have the rank test score is basically the sum of the square of all the differences in rank. So the difference in rank here would be 1 minus 4. It'd be 3 squared, and you take the sum of all the squares and you plug it into this S which is going to go into a correlation coefficient.

Similar to in values and hypothesis to the Pearson, but now not making assumptions about the parametric, you just talking about ranks, it's a non-parametric test. And then, the N is the total number, and you apply that formula. Now, they apply that formula to this exact roughly the same data set or very similar. Actually data set.

Another critique they had was that using a better measure of the protein abundance would be using a radioactive tracer in the protein. And then measuring quantitatively the intensity of beta particles released from the methionine in the proteins.

Making that change and comparing it to the same messenger RNA assessment, they got this fairly linear correlation over three logs. Using the Pearson correlation coefficient, which they didn't entirely approve of, they got a 0.76, which is a modest-- it's a significant linear trend. And using their rank method 0.74, which is basically very similar.

And found no significant difference between the top 33% and the bottom 33 proteins. Undermining the previous claim that there was less linearity at one than in another. You might, or you expect that the least abundant majority of proteins would have a little bit of either biological or instrumental noise. Nevertheless, this group found that it was a good correlation.

Now, these two plot-- the plot in the next slide looks similar, but it's really quite different. Here the Y-axis remains. The protein abundance is measured by this F 35 labeling, but now we're getting back to this game of asking to what extent can we make predictions about the properties of the proteins? In this case, their abundance is based on their use of abundant codons.

And a way of quantitating this is coadaptation index. Shown in the lower left part of the slide is you have-- it's the log of the coadaptation index, is a sum of all the frequencies F of I of each of the codons. Where I is the 61 non-stop codons out of 64 total. And

The W sub I is a waiting factor, where it's the ratio of the frequency of codon I. Let's say a leucine codon there are six different leucine codons, and say the first one, the W sub 1 is going to be the ratio of the frequency of that first codon to whichever one happens to be most abundant one. Could be the first one, or it could be one or the other six. And so, that's the formula that is used.

And you can see, again, a nice linear trend where indeed, the most abundant proteins do tend to use the most abundant codons. The codons that you find that are most abundant both in transponder RNA level and in usage in abundant proteins. Now, just as with RNAs, you can measure them on an absolute scale or a ratio scale. The advantage in principle if you have it on an absolute scale, you can always calculate ratios from it.

But not necessarily vice versa. You have ratios you can't always get to absolute. So that's one advantage to absolute. If you're looking at things like codon adaptation index or messenger RNA, or a variety of other motivations for you, need to do it on the absolute scale. But there, an argument can be made for doing it on a ratiometric or relative scale, in the sense that you can establish internal standards which are more precise.

You can really eliminate many of the systematic errors that can creep in due to differentiable ionization in the case of mass spectrometry, and so on. Now with RNA, the way we did this ratio testing ratio quantitation is we would label one messenger RNA red or site 5, and the other one green, and then by using selective filtration in the imaging, you could get ratios.

With mass spectrometry, you don't have colors. But somehow, you want to get the same idea across and so what you do is have masses. So what you want to do is encode them in something that will not change the chemistry, but will change the mass. Not change the mass too much because if you change too much, you might change the chemistry, or you might not be able to find the shadow peak, the second peak.

So basically, what we're doing is we have a-- want to do is take self-state 1, label it with a light ICAT reagent, self-state two, label it with a heavy, meaning more neutrons in it and then mix them together as early as possible. And then all these steps that could have little systematic errors in it will be the errors will be equally distributed to each of the test sizes and labeled and then measure the mass.

And for each mass, you'll have two peaks that are separated by whatever the difference in mass between the labeling changes. So what are the properties that you want of a labeling agent? One is that it should react covalently so it will survive all these fractionation steps and mass detection steps. So it has to be, in this case, file specific. You want a way of pulling out only those peptides that have been modified. All the rest are just going to contaminate your mass spectrometry.

And in between, you want something where you can differentially label heavy or light atoms. In the original proof of concept, this was done with hydrogen atoms versus deuterium. They differ by one atomic mass unit per position. However, you will see in the next slide, this has the unfortunate consequence that hydrogen and deuterium actually are not only mass distinguishable, but they also have different chemical properties.

There's an isotope effect that's detectable in a variety of chemistries, including retention time on the HPLC. This has since been upgraded for you having C13 versus C12. Now, that's a much more subtle difference. Same mass difference. You have nine different carbon atoms in here, and that works better. In addition, there's a way that you can cleave off the biase after it's done its job, where you would add it, and you selected all those peptides that have been modified and then acid cleaved all that to clean up the mass spec.

So here's an example where you have the difference in M over Z, of 4, M over Z and it's between these heavy and light peaks. And the ratio of these two is something you can use for essentially every peptide where the pair of peaks is in an uncluttered part of the mass spectrum. And here's evidence where you can see that on retention time and the horizontal axis on the left-hand side of Slide 41.

You can see this little red line shows how the centroids of the set of peaks should line up and has been displaced through the chemical effects of that adding a few neutrons. OK, so the lesson here is that hydrogen deuterium are not necessarily chemically identical. The conceit of isotopes is that they really should be chemically identical, but really it's better if you work with heavier atoms to introduce neutrons.

Now, so what can we do with this ratiometric acid? Now, we're going to-- just like before, we compared absolute protein levels to absolute RNA levels, now we're going to measure ratios. And to do ratios, we have to have two different conditions. So another advantage to absolute is you can do it all under one condition. Here the two conditions that were chosen for this proof of concept experiment were glucose and galactose.

That is to say, growing it plus or minus galactose. Galactose is a nicely understood metabolic and regulatory system in yeast. It fits in with what we know about the central carbon metabolism, and it induces a set of genes in blue here that are required for galactose catabolism to produce energy. The most strongly ones are way off in the upper right-hand corner here, Dow 710, and one, these are the core catabolic enzymes.

But almost all of the innate blue triangles have some kind of story like that. And these all are in the upper right quadrant have a high log 10 ratio of expression up to three logs to the third fold induction. Similarly, at the other end of the spectrum in the lower left-hand quadrant are respiratory genes that are involved in, say, oxidative phosphorylation.

And these are moderately depressed under galactose conditions, and that's why they're in the lower right-hand corner. And then the green ones are not quite along this diagonal. Their messenger RNA is increased, but their protein expression is not. And these are the ribosome protein genes, and this is another phenomenon that's well documented in the system.

OK. Now those are examples of how you can use absolute and relative, and why you're motivated to use absolute and relative measures for proteins and messenger RNAs. Now, these are all treating as if, just like before, I said that messenger RNA might lump all together all splice forms and call that the gene product. Although you know better, and the same thing with proteins, you might lump together all the protein splice forms.

And not only that but for a particular protein splice form, there are many different synthetic modifications, such as proteolysis and phosphorylation. And so we're going to talk about these modifications very briefly to hopefully whet your appetite for one of the most exciting parts of proteomics and whether it's identification or quantitation. So we've already mentioned radial isotopic labeling as a way of quantitating.

Using various radioactive sulfur, you can use, whether it's stable isotopes or radioactive isotopes, you can use these to do pulse labeling to monitor a dynamic process. P-32 if in particular, if you want to enrich for some of the most well-studied and significant photosynthetic modifications involved in signal transduction. You can enrich for particular types of amino acids.

We already showed that the cysteines, which is an arbitrary amino acid chosen for the ICAT ratiometric method, that was chosen because it has interesting reactivity, not because it's intrinsically important low abundance regulatory molecules. Phosphates, on the other hand, can be very important, and you might need to enrich because these important regulatory phosphorylation sites can get lost in the snow of all the rest peptides in the proteome.

This can either be done-- this enrichment can either be done by immobilized metals such as iron and gallium, and so forth. This is called imac for immobolized metal affinity chromatography, or you can have antibodies that are specific particular phosphate peptide, and particular phosphate amino acids. You have lectins for carbohydrates as front ends for mass spectrometry.

Even when we do P-32 labeling metabolically where the P-32 will only label the subset of the proteins which are phosphorylated. It is still the case that some of the most interesting regulatory cell cycle proteins are not detected above background because there are many abundant proteins, such as ribosome proteins, central carbon metabolic enzymes, which are needed in high abundance, but also need to have phosphorylation.

And so you get this forest of phosphate proteins such as ribosome and metabolic, which make it hard to detect the regulatory ones. So labeling is not a panacea, and we'll come-- I think you'll see as we go through the protein modifications and mass spectrometry, is a multidimensional purification really is the way that you get away from the ribosomal proteins and the highly abundant metabolic proteins, which are interesting, but you need a way of both setting them and the regulatory low level proteins.

Here are some examples of natural processes. So you can think of this as a special class of post-synthetic modification. Instead of having a phosphate glomming on, you have two different peptides either intramolecularly within a protein or intermolecularly between proteins. And you should be highly motivated to study these because they tell you something not only about protein structure, three-dimensional structure, which was the topic last time, but protein-protein interactions.

And not just theoretically, what proteins might interact with other proteins, or might bind in vitro could be in vitro artifact or in yeast to hybrid system, could be a yeast to hybrid artifact. These are actual covalent caught in the act in vivo protein interactions. And some of these are-- most of these are very well documented.

By far the most common one and of great significance to protein tertiary stability in the class of proteins, which are extracellular. These include extracellular domains of membrane proteins and secreted proteins, and, in particular, because there the oxidation state is such that this sulfur sulfur bond is stable, while intracellular tends to be more reducing atmosphere, and so that these disulfides have trouble forming.

Collagen has a lysine cross-link. Ubiquitin has a C terminus to lysine cross-link. Fibrin involved in blood clotting has glutamine glycine, and so on. As some of your proteins age, you will find glucose in high concentrations in your blood will glycolate the lysine residues. And this is part of the process by which these proteins eventually lose their function and are cleared.

Protein nucleic acid interactions we've been talking about so far are non-covalent. Some of them are covalent, for example, when you want to prime de Novo polymer synthesis, DNA synthesis. OK, so what are the consequences for the mass spec algorithms we've been talking about, say, de Novo sequencing or finding a peptide spectrum in your database? Well, you can see the masses are going to be fairly straightforward.

Here's some examples of some masses, of some peptides, and some cross-link peptides. On the top right, you'll see one intermolecular cross-linked between a lysine and a lysine, and an intermolecular between two peptides. Now each of these peptides in this display, these are just simple masses. These are not fragments. These are not some fragments. So you expect these to be triptych products.

So each of their C terminals should be either arginine R or lysine K. R, K, K, R, so forth, and so you can see this is two peptides, one ending in R, one ending in K. So let's look in detail at this example where you have an intramolecular cross-link. And you can see that as you cleave it, each of these peptide bonds entering B ions from the N terminus and Y ions from the C terminus.

You'll see that there's a special case in the region that's defined between the two cross-links and that any peptide bond cleavage that might occur in the gas phase when colliding with argon or some other inert gas will break the chain as usual. But the chain won't fall apart because it's got actually two connections. One is through the normal peptide bond. And the other is through the cross-link.

So cleavage is all through in here. It takes two hits to get separation of these, and two hits is unlikely. And so, you'll tend to see the B ions in the Y ions right up until you hit the first cross-linked amino acid, and then you lose it. So that's one of the complications that you have from cross-links. The other one that when you have, say, cross-linking two peptides, might occur when you have an interaction between different proteins, is you'll now have two sets of B ions and two sets of Y ions.

As If just having B and Y in the same spectrum isn't enough, now you've got two of each, and even though you don't have the cycle to worry about, you have a full set. But there is a algorithm that Tim Shen and others have developed for dealing with that in very clean cases. And here's an example of actually using the cross-links that you get from mass spectrometry as a fairly inexpensive set of constraints that you can use for getting distances either intramolecularly in this case or conceivably intermolecular.

And the constraints. You can't get any bigger than the cross-linked distance. You know that the chemical structure of the cross-linker, and so you can say these two amino acids with their side chains and so forth are reacting have to be this link or shorter. And that's what these little yellow lines indicate in this fibroblast growth factor two, FGF two, where the crystal structure of FGF two is known.

And these constraints will greatly aid your ability to find distant homologs or to increase the precision of your homology modeling in three dimensions. The shorter you cross-link, obviously, the better, the tighter your constraints, but it might reduce the efficiency of cross-linking. If you're doing this as artificial cross-linking as opposed to if you have a natural cross-linking, you're basically stuck with whatever the natural.

This was an artificial cross-linking with a chemical cross-linker, or five functional cross-linker. That's just a reminder. This is a different way of showing some of the things. We had a scatterplot last class, which showed that as you increase the sequence identity in homology modeling up to 100% on the vertical axis, you decrease your uncertainty and the observed root mean square deviation that you get between two structures.

But they're better than 80%. Then you have in the order of one Angstrom deviation, which is quite acceptable for many purposes. Now, if you want to do threading to very distantly related structures getting down around 30% is getting at 25% is getting the Twilight Zone, where you really can't believe it. It's off by too many angstroms.

But these constraints in the previous slide could help you out in either doing the mozzie modeling or doing a threading where you're searching through your favorite sequence through a database of three-dimensional structures to ask which three-dimensional structure is closest to. Now that FGF two, we had two slides back with all these constraints can be run through the threading algorithm, where you run the sequence through the database illustrated here to this horizontal, then to the various rows of different structures here.

The fold family is in the second column from the left. The sequence identity of our search sequence, which is FGF two, against all these three-dimensional structures that present identity here 98.6 is basically, that's the same structure. That's the trivial example because threading rank is number one, as it should be since it's exactly the same structure-- almost exactly the same structure.

The constraint area, of course, is going to be zero because all three-dimensional structure, and all the cross-links work to that structure. But an interesting one now this has been ranked by the constraint error, not by the threading rank. And so you can ask does this improve the hits? And the next one down is FGF two compared to aisle one data, and they do have the same fold family.

We know that from three-dimensional structure, and the percent identity is way below the usual cutoff where you can't infer from threading or sequencing. In fact, the threading rank is five. It's not the second-best thread. But it's a straight arrow zero, and so if you combine the good threading rank in the constraint error, then you would put this as your best just homolog 12% to 13% sequence identity.

And of course, it's beating out better threading ranks because it has fewer constraint errors. So you can see how powerful these constraints might be, and it's certainly-- you just need to evaluate just exactly how cost effective the mass spectrometry is. Now last topic today, in the realm of protein modifications and interactions, are how we quantitate metabolites.

Now you can see that we've got some momentum here on quantitative proteins in RNAs. And so what are the issues that are slightly different from metabolites? And Slide 52 summarizes some of these. You have when you break open a cell to isolate messenger RNA or proteins. There is the rate at which degradative enzymes act is on the order of seconds.

That's the rate at which they go, while many of other metabolic processes take on the order of milliseconds to microseconds. Very rapid kinetics, and so as the cell starts to get a little bit sick on the second range, all these enzymes are scrambling the metabolites concentrations. So you have these rapid changes. The detection methods are historically idiosyncratic. They might be enzyme linked, where you'll have in order to detect the metabolite, you have a series of enzymes that result in some fluorescent or luminescent assay.

Or they could be gas chromatography, liquid chromatography, NMR mass spectrometer, and so forth. The good news is there are usually fewer metabolites than there are RNAs and proteins. There could be 30,000, some RNAs and proteins typically only 1,000 or so metabolites, even in the more exotic, the metabolically enabled such as E Coli.

Here, from their various databases, ecosite, width, tag, and so on, which integrate information about metabolites with the enzymes to act upon them. Here, we're just looking at the mass range that we have. Typical mass range, they're very small compared to proteins in RNAs. Most of them being around 200 atomic mass units.

And many of them having absolutely identical mass, that is to say, they have atom per atom exactly the same composition, even though it's arranged in three dimensions very differently. For example, isoleucine and leucine, as their names might imply, have exactly the same mass no matter how many significant digits you put on them. And this is illustrated by actual data on isoleucine and leucine.

These supposedly highly purified versions commercially available of isoleucine and leucine mixed together here and run out in these two dimensions of mass and the horizontal axis and retention time and hydrophobic separation. Now, not a peptides but amino acids, metabolites, and you can see how that even though they are identical in mass, as shown on the previous slide around 131, they are separable by their hydrophobicity.

They have the same atomic composition, but they are separable just by this hydrophobic separation. And you can see in the commercial press, there are a variety of contaminating molecules that co-migrate in their reverse-phase dimension, presumably because something like reverse phase is used for purifying them commercially.

So there are basically three ways of distinguishing molecules that have the same mass. The one in the previous slide was separating them by another property, like retention time on hydrophobic properties. Another one is secondary fragmentation. Just as we could fragment peptides by collision in the gas phase with some inert gas, we can do this with metabolites.

And so two things that have the same mass may have a different fragmentation pattern. And you can see again, you can cleave it every particular position, and here are two different aspects and two different labs. Slightly different methodology showing fragmentation in almost every carbon bond. The third method by which you can distinguish compounds that have exactly the same mass, and this case this is the most extreme case.

These compounds have the same mass. They actually not only have the same chemical composition, the same atomic composition, they actually have the same chemical structure. Their three dimensions are the same. Their mass is the same. What it is is this-- let's say, the red is a carbon-13, and the green are the carbon-12. You can have the carbon-13 in different positions on this, say, this glucose molecule.

So it has the same four-dimensional structure, the same mass. It's just you moved the position of the C-13 to different positions. This is actually an interesting case when you have natural abundance glucose, where you have C-13 trace amounts, it can be positioned on various different carbon atoms. But you can still tell where it is, by when it-- it's now not broken down in the gas phase like collision-induced dissociation.

But it's broken down in the cell. It goes through different pathways. And this is the example-- that we're going to be talking about pathways non-stop for the next three sessions. But here's an example of central carbon metabolism, where you start with glucose and glucose phosphate in the upper left-hand corner of this network diagram, and you end up with carbon dioxide down on the lower left. And it can go through various pathways through ribulose or down through three carbons.

And each of these three and two carbon breakdown products can have the labeled atom, the mass tagged atom in different positions. And as this quote from literature points out that when you want to study the fluxes through the pathway, by monitoring, you can actually do a pulse or a stable steady-state labeling with isotopic labels. And you can monitor the fluxes through these pathways.

But you need to take into account all the different ways that you can go through the pathways. Especially, when you're doing metabolic cycles like the TCA cycle or you have to think through all the multiple turns. Now, in principle, that kind of metabolic tracing can be done either with mass spectrometry or with nuclear magnetic resonance.

When you do quantitative 2D nuclear magnetic resonance, you're basically looking at the shifts in the spectral quantities for the carbon-13s. Remember, we were talking about carbon-13 labeling and the normal, most abundant protons. The chemical shifts here that you get are due to the exact chemical environment of this proton or the carbon-13 for, say, the alpha for each of these amino acids, alpha betas.

And each of these little clusters are schematic for the intensity of these particular atoms that you're monitoring by their isotope effects on the NMR here. The odd number of nucleons is critical to the detection. OK, so if you know the structure of the network, then you can use that knowledge, and you know which of these atoms go into which parts, then you can use this to quantitate the fluxes through any point in the network.

On the other hand, if you only know part of the network, then you can use this way as a tracking to slowly piece together how the network must go. Most of this is both worked out well before genomics and our current systems Biology methods, and so there aren't real algorithms for doing this as far as I know. Although, certainly, there's an opportunity for doing it.

Now, this is measuring-- remember this ratios versus absolute amounts. This is measuring not only ratios of metabolite concentrations. These are not metabolite concentrations, but fluxes. So with metabolites, you can measure concentrations or fluxes absolute or ratios. All four of those combinations. Now, let's say-- again, in principle, if you can measure absolute concentrations, you can measure ratios, and you can measure ratios of fluxes.

So how will you measure absolute concentration? So remember, we said that one of the problems was that as soon as you start perturbing the cell in microseconds, you can get changes. But one way to do this is without slicing the cells. You can snap freeze them first. Yes, snap expose them to aqueous methanol at -40. You can wash them at that, it's a liquid, and you can remove the outside metabolites.

And there's reason to believe that this is the minimally perturbing method of preparing the cells. And then you put them in basically boiling alcohol, and then quantitate with NMR methods such as the ones we just talked about, getting up to 1,300 measures per sample. Some examples of some of these internal metabolite concentration, now, remember, these are not flux ratios.

But actual metabolite concentrations of things like glucose phosphate, ATP, pyruvate, so on, can be correlated to genomics by the vehicle of gene knockouts. We have wild type on the top of the far left, followed by a deletion of HO. This is a homing in the nucleus, should have no metabolic consequences at all. This is used as a control. It's a pseudo wild type, just shows that the deletion method itself is not changing things, the metabolism.

And then, as you go further and further down this list, you get more and more severe expected effects on the ability of the organism, say to produce energy, as exemplified by its ATP production. Now, if you have low ATP production, that means you have high residual levels of ATP, which is the other end of the energy spectrum. And so if you look down these columns, it's a little hard to see with all the clutter that's produced by the standard deviations.

I wouldn't want them to get rid of those standard deviations. It's wonderful, but anyway, you can summarize this by looking at the ATP to ADP ratio. That's a way you don't need ratios for this to work. But it's a way of accentuating this energy balance, and so you can see, for a wild type, you have about almost 7 is the ratio of ATP to ADP. Highly charged on this high energy state.

And as you get to these pet mutants, which are mutations in the mitochondrial process, you find the cell is becoming increasingly ineffective. Now, these were all chosen because these were so-called silent mutations. The title of this paper says that you can get phenotype by looking at the molecular analysis quantitating the whole holistically, systematically all the metabolites for things that otherwise have no phenotypes. OK.

Now, this representa-- as you start quantitating whole proteomics, proteome interactions, proteome modifications, metabolites in their interactions, you start wanting to relay this information. Summarize this information in the context of models. And just as last time, the upper left-hand portion of this summary, I emphasize that no model is exact.

Sometimes the people working in the field convince themselves that it's more exact than lower models. But every one of these, even the quantum mechanics, is a poor approximation of electrodynamics, and molecular mechanics has only spherical atoms represented. Some of the most challenging cell models involve massive equations of stochastic. Now, we're not talking about single atoms anymore. We're talking about single molecules. Still it's too coarse for many experiments.

You can have phenomenological rates, such as the ones we've been talking about, represented in ordinary differentiable equations. How do you get the parameters that describe now concentration and time? Not single molecules, but treating them as a bag of a particular part of the cell, or the whole cell having a particular concentration of molecules. And then we'll go on as we get into the network analysis, and the next few lectures, we'll talk about some of these other models, which have their roles.

But let's just get at-- how would we get some of the parameters that describe concentration in time? And when we talk about the formalism of these networks, regulatory networks, mainly are about binding. But they intimately connect with catalytic networks, where you not only bind, but you actually change the covalent structure of molecules. In the simplest such case, single substrate going in single product coming out is that the top of this Slide number 60.

And here, you can see that the enzyme E, which is typically a protein and/or but there are RNA catalysts and so on. But the property that's emphasized here is that the enzyme is not consumed. As A goes in, it makes a covalent change from EA to EB, B is released. E is recycled. E is not consumed by the cycle, but A is consumed. B is produced.

But let's look at it in a different light. In a particularly interesting class of reactions, for example, those involving the regulatory cascade signal transduction enzyme modifications. Here the enzyme now becomes substrate. The ATP is now no longer consumed in the whole cycle. ATP comes in, modifies the enzyme to a phosphate enzyme.

And the accompanying reaction regenerates the-- turns the ATP back into ATP. So the ATP, in a certain sense, is catalytic here, and now the enzyme is consumed, producing a fossil enzyme. So it all depends on when you started constructing these graphs. What is the node-- within the node is a substrate, and node is an enzyme, depends on how you look at it?

And I do this somewhat provocatively, so you'll think about these networks not just as binding, but as catalysis. And not just the enzyme being the catalyst, but sometimes the substrate as well. Now, this is the simplest case of measuring kinetics. This is not equilibrium. This is kinetics. And we're studying this plot on the far left-hand side of Slide 61. As substrate increases, the rate of production of product increases.

You could start out with zero product or small amounts, and it used to be historically at bottom, you would require to do the experiment, you would require that the product be as close to zero as possible. You'd do initial rates, and you have this simple relationship where the increase in product was a simple function of the 1 over 1 plus reciprocal of the substrate concentration.

And as the substrate would increase, you'd eventually saturate the amount of enzyme that you have present in the experiment, and that would be the maximum velocity for that amount of enzyme would be max. However, if you have a more full equation where you take into account all the players, at least in the simplest system, the forward velocity, the PDT the derivative of product concentration as a subject of time, this might be in mol/Ls per liter.

It's going to go up as a substrate goes up. More substrate means it will go faster in the forward direction towards product. But if you have some product, you'll have some product inhibition. You'll have a tendency the kinetics to go in the opposite direction. So this is negative-- there's a negative component with products. The KS and KP are sometimes called Nikolaus constants as the substrate gets closer to the Nikolaus constant. This is basically related to the binding affinity for the substrate.

And that's a natural halfway point where the substrate is half saturating is roughly what the KS is all about. Now let's compare this very simple case. We have one substrate producing one product to a more typical case where you might have two substrates producing two products. And let's take this out of a real network. We're going to show this real network in just a moment, and it's all slurry.

Here you have two substrates ATP, and F16 going to 2 products ATP, and FTP. And you've got this same form where you have a velocity of this reaction. And it's a function of the reactants F16 and ATP, and you find them in the numerator here. And you find these Nikolaus constants in the right places. But in addition, you find in the denominator this curious term, it has these fourth powers.

Well, we didn't see any fourth powers in the previous slide. Why are we getting fourth powers just because, and why is AMP in here? It's not even one of the reactions or the products. What's going on? Well, this is actually a regulatory phenomenon allosteric where you have a second site on the enzyme. One site does the catalytic magic, and the other one is regulated by some in the infinite wisdom of the whole network, that is important feedback.

So AMP is related to ATP as a further step, and it feeds back on this enzyme, as does F16, and ADP, and so forth. So all these-- and this fourth power just says you want it to be cooperative. You want to have a nonlinear regulation. That's what's going on in this term. And it doesn't occur in the whole network, we'll show in the next slide. But it does occur at some key points, like this one where the enzyme has two sites, catalytic and regulatory.

And when you see these terms that are greater than linear, like the fourth power, that's sometimes referred to the Hill coefficient and refers to the steepness of-- instead of having this kind of curve, you have this kind of sigmoid curve. And the sigs peakedness to that is related to that power. Now, let's look-- if you compare-- so this little piece, this phosphofructokinase step, is going to be put in context of the entire network in the red blood cell, human red blood cell right here in the upper left quadrant of the circle.

And this is the simplest analysis-- this is the simplest metabolic network that you'll be talking about. This mostly involves covalent transitions or pumping across membranes. It treats the whole cell as a uniform bag, with a membrane being a separate compartment. And there are really two objectives of this. One is to produce ATP. So that you can run the pumps to keep the osmotic pressure constant across the red blood cell membrane, so that it maintains its shape.

And the other is to maintain the redox at the right level, so as if-- so you have a reducing atmosphere. The hemoglobin has an intimate contact with oxygen, and so there is a certain low level of oxidation of the Iron rather than just binding to the Iron to make hemoglobin which is not a good physiological state. So you want to have this reducing potential to get it back to the correct oxidation state, so it will bind oxygen.

So you have a little bit of purine metabolism here. Quite a bit of simple glycolysis, not oxidative phosphorylation. Just because glycolysis produced reducing potential and little ATP, and these 40 or so enzymatic reactions can be modeled with about 200 parameters. All of these 200 parameters have been measured accurately by purifying each of the metabolites and enzymes.

And this has been reduced to an ordinary differentiable equation model that has been evolving since the '70s. If we look at Slide 64 that this phosphofructokinase, we have the same form that we had a couple of slides back. Here is the fourth power term, and the denominator having AMP as a regulatory molecule. And then you have the Nikolous constants, and so forth, explicitly stated here in the numerator for the substrates F16 and ATP.

And you have a similar equation for every single enzyme step in that whole network in the red blood cell model. And in green are the concentrations, at any given time point, that's going to be their dynamic or steady state. At any given time point, you'll have green for the metabolic concentrations and red for the fluxes indicated by the enzyme. And you can run this as a simulation for the red blood cell and see all the interesting questions about robustness and optimality, and so forth.

Even though this is a very mature model, there still is lots to be done, even in this very simplest system. Now, we're going to go on to more complicated systems in the next three lectures on networks. But for today, we've basically tried to integrate the protein, either absolute or ratios, with RNA measurements and with metabolite interactions, proteins post-synthetic modifications. So until next time, thank you very much.