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.

GEORGE CHURCH: OK. Welcome to the second proteins. Just as with the RNA analyses, we started with, in that case, a brief discussion of RNA structure, and then moved on to RNA quantitation. This case, we spent a bit more time on protein structure, and we'll spend a little less time on protein quantitation, which will be the main topic today. But partly because the RNA and protein quantitation have many themes in common, so we've covered some of those. And we'll talk about how one can integrate protein quantitation with RNA quantitation and with metabolism.

So last time we talked about mainly with structure and interaction of proteins with small molecules at the structural level. And this time we'll be building up towards the quantitating proteins and their interactions with other proteins of small molecules so that we can address that as the first step towards the network analysis that will be the subject of the last three lectures of the course. So we'll get a hint of that at the end of today's talk.

So at the beginning of the course, I had one slide talking about how purification was a revolution in many different fields-- one of them resulting in recombinant DNA and the Genome Project, and so on. Today we're going to talk more about purification from the standpoint from what does it tell us about the properties of the molecule itself as opposed to just purifying for its own sake. How can we go in and data mine and get the maximum value out of purification?

The things that we expect from purification is to reduce some source of noise. That is to say, in the process of identification or quantitation of proteins or their components, we want to remove sources-- the major source of noise would be environmental contamination, or more often contamination with other bona fide members of a mixture, other proteins that may be very abundant or breakdown products or related products. And we want to separate it into enough different components so that the major ones are off in their own little bin and don't interfere too much.

So the second reason why we might want to purify is to prepare materials for in vitro experiments. While we're studying networks in the next three classes, we'll find that the real problem is that they're complicated enough that you need some way of isolating one network component from all the possible interactions. And one way to do that is to purify that network component, or a number of network components out and make a subnetwork completely artificially in vitro. So that's another use of purification.

And finally-- and this is the main theme for the next couple of slides-- is to discover biochemical properties about that component itself. So what can you glean from the purification process? And this requires careful use of purification. Now, many of these methods were developed because they work well, but now we go back and we look at which ones can give us information about the biology and the chemistry of the system. And so we have the charge of a molecule, which we talked about last time as being a very important relatively long range interaction 1/r interaction.

And here are two methods I mentioned, one involving electric field, isoelectric focusing, which determines the pH at which the charge is neutral and there's no net movement. In ion exchange chromatography, we have a mobile phase and a solid phase. And charge enters prominently into that. Size is something that will be a recurring theme tonight, where we'll be talking about the mass of protein complexes, the mass of individual protein subunits, and the mass of peptides cleaved out of those basic subunits.

And you can see there are quite a number of different methods. Sedimentation velocity will be one that we will use-- electrophoresis, and so on. Solubility and hydrophobicity I'll lump together here as properties, kind of bulk properties of the amino acid composition of peptides and proteins. And they refer to their affinity for hydrophobic solid phases or hydrophobic mobile phases, solvents, and so on. The biological significance of hydrophobicity might be the affinity for lipid bilayers or affinity for other hydrophobic patches of other proteins.

Now, we have-- the size tells us something about the stoichiometry of protein-protein interactions when we're looking at native measures of size of complexes. And other indications of specific binding, whether it's between proteins or a protein in a small molecule can be detected by affinity chromatography or another related method as [INAUDIBLE] precipitation, where you'll have one ligand, like an antibody, which is specific for a particular protein epitope and will pull down all the proteins that are associated with that epitope, that surface property.

And another sedimentation method-- now, the first sedimentation method was a velocity method, a kinetic method, where the largest particles, once you take into account buoyant density, all other things being equal, the largest particles will sediment most quickly. If you set it up so that things are nearly equally buoyant and you build up a density gradient by centrifugal field, then you get particles separating by their properties, which can include the binding of metal ions, which greatly affect the density of nucleic acids, for example, and hence, nucleic acid protein complexes.

Now, this is a particularly awesome pair. And historically, it figures prominently into proteomics. It's still quite viable. And they illustrate some important points. When we talk about the mass of a native protein or of a protein that's been denatured by a detergent micelle, such as Sodium Dodecyl Sulfate, SDS, its mass can be resolved by fairly potent polyacrylamide gel electrophoresis, where you get sieving, where the gel causes sieving. And this micelle, this detergent micelle size itself, or the micelle is dependent upon the size of the unfolded or partially unfolded protein chain. And so you actually get a fairly good calibratable plot of mass of the protein versus the mobility in an electric field when the protein embeds itself in this detergent micelle.

So this observation of the potentability of this detergent to denature proteins, and then to resolve based on mass with the small exceptions of very hydrophobic proteins or very carbohydrate-rich proteins-- most other proteins will have a very nice calibratable relationship. Similarly, the charge on a protein approaches 0 as the pH gets to the point where it's titrating out all the type titratable groups. And this can be calculated. And the resolution of this is about 1 part in 100 or better. So both of these together are two very high resolution methods.

And you can think of trying to divide-- you have a complex protein mixture. You want to divide it up into a lot of separate bins, maybe 100 of each. And if you combine two dimensions, as you often do here, where you'll run first isoelectric, and then second the SDS dimension, now you got 100 bins in one dimension, and each of those bins turns into 100 bins in the second dimension. So theoretically, you've got on the order of 10 to the fourth bins. And if some of the things that you want to-- some of the rare proteins you want to analyze are in one of those bins, you might have gotten as much as a 10 to the fourth-fold enrichment for those rare proteins away from the more common ones, which are all too easy to stumble upon.

Now, before we get to an actual example over two-dimensional gel, I want to motivate this by the computational components to it and the computational biology that you can get from these multi-dimensional separations. And this comes from our recurring interest in comparing whenever we can calculate a property of a system, or in this case of a protein, we do so, and we compare it with the observations. And some of the properties here are the localization in the cell, which essentially is an association of the protein as it's made, the post-synthetic modifications to the proteins, such as proteolysis and phosphorylation, and so on, its charge, and its mass.

And here is shown a plot of calculated charge, or isoelectric point in pH units. P being the negative logarithm of the hydrogen ion concentration. So it's calculated on the horizontal axis and observed on the vertical axis. And what we're seeing here is if there were a perfect x equal y relationship here, where calculated and observed were the same, then all the dots would lie on this line, and though the outliers, initially they were such things as frame shifts of the DNA sequence, producing a wrong calculated protein. Once those were corrected, then there were observed proteolytic cleavages, which could be mapped down through the exact amino acid, and then those corrected a few more on to the line. And the remainders were other post-synthetic modifications, such as phosphorylation.

So here's a reason to embrace your outliers. Each one of these things is an exciting story, where you either have a correction to a previous data type or a new discovery of a post-synthetic modification. Now, how do we actually calculate all these facts about protein properties? Some of these have biological significance, which we've listed. Where it is in the cell, obviously, matters to its carrying out its function. How big it is determines what other proteins are associated, and so on. So how do we calculate this?

The protein charge which was in the previous slide is a simple linear function where you sum up the pKas of each of the individual amino acids. Now, pKa, again, p means negative logarithm, and the Ka means the equilibrium association constant of the proton with each of these chargeable residues. So depending on the pH, there are a wide variety of nearly physiological pHs where [INAUDIBLE] and histidines will be positively charged, the blues, and the red ones can be tyrosine and cysteine. Especially aspartate and glutamate can be negatively charged in the range of pHs that we saw in the previous slide.

And so this is calculated as that simple sum. And protein mass is calibratable with knowns. Even if you have a very complex empirical relationship, such as that detergent binding FBS gel electrophoresis. It sounds very-- too many moving parts to be completely theoretical. But if you calibrate it with good known proteins, or protein complexes if you're doing a native electrophoresis or native sedimentation velocity, then you can get a curve where you can interpolate and find masses quite accurately, or at least around about 2%.

But mass spectrometry is commonly applied to peptide masses, sometimes whole protein masses. And here, assuming the mass spectrometer is properly calibrated and so forth, this is a simple isotope sum. And this can be carried out sometimes to four or six significant figures. And it really is a simple sum of the isotopes measured by physics. And this can include post-synthetic modifications. As you're getting down to peptides, the post-synthetic modification becomes a much larger fractional effect on the measures that you're making.

Not only the mass, but the liquid chromatography properties of either a protein or a peptide can be calculated. Here you take amino acid composition and do linear regression on a calibration set. And you can get precision on the order of 5% or better.

A subset of localization, we have motifs. Sometimes hydrophobicity is a part of those motifs in their description. Expression would not be something-- we've seen the motifs that are involved in regulation of transcription and so forth. But kind of the shortcut that might take you directly there in certain cases is the codon adaptation index. This is something where the hypothesis is that you can go directly from the nucleotide sequence of a protein coding region, and if it uses codons that are very abundant, that correspond to very abundant transfer RNAs, then that protein-- then that's saying that the evolutionary pressures producing that particular choice of codons is revealing that that protein is going to be high abundance. So in a way, this is a way of going directly without-- due to this observation and somewhat logical expectation. so

Now we have all sorts of separation methods and motivation for studying them, more than just using them. But now we want to look at a particular case of separation. We're talking about complexes and protein localization.

So this here is another example of a two-dimensional gel, where the isoelectric point is on the x-axis here and the vertical axis is this estimate of molecular mass, provided by the association of the protein with the SDS micelles and the effect on gel electrophoresis. And this is a 2D gel, a small section of it, not the full pH range nor the full molecular weight range, but this has a good fraction of the proteins that are secreted.

Now, to what extent can we calculate this? We've shown a number of calculations and observations so far. And we pointed out in last class that to some extent the transmembrane regions of a protein could-- this is one of the better algorithms in protein sequence gazing.

And taking that one step further, you can actually say, OK, we know that this might have a motif or two that interact with membranes, or somehow are part of the process by which proteins are targeted and move across membranes so that you can divide it into several different subcellular localizations-- in eukaryotes, the mitochondrial localization, and chloroplasts in plants, secreted proteins that go all the way across the plasma membrane, and other locations, such as within the membrane. And this can have an 85% success rate, meaning a 15% false negative. And you can see modest false positives here, too, over predictions in 295 transmembranes.

So let's return to mass, and this time in the context of mass spectrometry. What do we have-- what's our starting point? Well, if you look on the far left side of slide 11, you'll see the simplest of the atoms, hydrogen, and you would expect this to have an atomic mass of 1.

Well, since carbon-12 is assumed to be precisely 12, it turns out that hydrogen is not precisely 1 when you actually measure it. And it's not even good enough for government standards. When you actually, say, add a CH2, it adds up to 14. And that is discriminatable from a nitrogen-14 with a good enough mass spectrometer. As it turns out, for most biochemical-- protein analyses, you don't depend on the sixth decimal point. You can get away, really, with-- or certainly don't depend on having the 10 to the minus three atomic mass units as your resolution.

But you do depend upon in a much-- in a kilodalton size peptide being able to get 1 part in 10 to the fourth, 1 atomic mass unit. Another big consideration here is that not only are these things not exactly integers, but in natural abundance, they're a mixture of the major isotope on the far left and the second and sometimes third and fourth stable isotopes, usually non-radioactive isotopes, which are present in nature.

And the most abundant in this particular list is C13. And it's most abundant in two senses. One is it's the highest fractional abundance of any of these elements. And secondly, carbon itself is very common in peptides. If you cleave your protein up with trypsin, which cleaves c-terminal to lysines and arginines, you'll get on the order of 10 or 20 of these peptides per protein. And they might be 10 amino acids long. And so they might have on the order of 40 carbons in them. So now the fractional abundance of C13 with 40 carbons is getting close to unity.

And we'll see an example of exactly how this plays out in terms of the multiple combinations of isotopes that can occur. Sulfur, on the other hand, has more stable isotopes. It has four different stable isotopes, but each one is a fairly small fraction. And the probability of having a sulfur in a given peptide is low. I mean, that's the probability of having one sulfur. The probability of having 40 sulfurs is vanishingly small.

So we've gone through calculating all this charged mass. Now we're going to do liquid chromatography, in particular hydrophobic measures. And this all falls under the heading of high-performance liquid chromatography, which is achieved under high pressure typically to get it to go rapidly. And so you'll have these little-- you'll digest your protein with trypsin. You get a series of these peptides, say, 10 amino acids or so.

And then they're injected in a liquid phase. They bind to the solid phase by the hydrophobic properties. You might have it. And then you'll have a readout, where, as a function of time, you get abundance where the peaks can be measured by mass or ion counting and so on. And these can be collected or simply run into a mass spectrometer.

There are going to be two phases-- the mobile phase and solid phase. Talking about the mobile phase first, the hydrophobic tendency of any given peptide is going to be some kind of related to the sum of the individual amino acid components. And you can either have a isocratic elution, where you have basically constant migration speed, no change in the content of the mobile phase, or you can have the mobile phase change its composition, say, for something that's almost entirely water at the beginning to something that has a high organic content, up to 40% or so acetonitrile, and so that at the end of the gradient, even the most hydrophobic peptides now have just as much affinity for the mobile phase as the solid phase, and they come flowing off.

The solid phase has a number of different options. The main one we've been talking about, implicitly at least, is the reverse phase, where you have the hydrophobic carbon-18 alkyl chains immobilized to a highly porous media in a column that can withstand high pressure. Or it can have differing polarities, size exclusion, the same sort of things that we were talking about with electrophoresis, except now the force is a non-electric field, but it's simply the pressure differentiable between the injection port and the output.

So here is a specific case worked out example of how you calculate the affinity of a peptide to a hydrophobic column. This is a C18 column, referring to the number of carbons in the alkyl chain. So you can see that this is like a kind of a lipid type phase. It's the sort of thing you might see in the middle of a lipid bilayer membrane.

And in this plot, you have relative retention time along the vertical axis. And the residue number refers to having short peptides cleaved, sort of walking along the protein. These have been synthesized so they march along the proteins, very analogous to how in an earlier class Rosetta made synthetic oligonucleotides marching along the human genome.

And this was done in order to calibrate to see how the composition, the amino acid composition of a peptide might affect its mobility and allow you to calculate it. And so what you end up with here is a somewhat intuitive way of summing the contributions of each of these amino acids. Now, remember, this is done under fairly acidic conditions. So the normally charged acid groups will now be protonated and be neutral.

And so what we'll find is mainly a spectrum from the very slowest to be eluted. They require the most organic, and hence the most-- the longest retention time would be the highly hydrophobic aromatics, tryptophan and phenylalanine, have a component which you obtain by linear regression of each of these peptides. You know the sequence of these peptides. You calculate their composition vector. And then you relate it to the retention time that's observed, so that the lower plot is the observed ones. And then after you do the regression, you get a set of these coefficients. You can now plug them in an additive sense to make this calculated plot. And you can see that there's very good correlation up and down.

It might have been better if these authors had done this as a plot of calculated versus observed, as we did in the previous one, and then showed the scatter and did a regression curve. But you get the idea.

And so the most hydrophobic ones are at the top, and the ones that are acid conditions are the most highly charged. There's the positive charge ones, the lysine, the histidine, and arginine down here at the bottom, and the acidic ones, since they're close to neutral or near 0 here. And there's a slight effect as to whether the amino acid is at the N-terminus or not.

So now we've calculated the reverse phase behavior here on the far left-hand side of slide 17. And this is separated by hydrophobicity, and then we have mass, which we've outlined how you can calculate the mass. So now you can calculate both its hydrophobicity on the vertical axis, and mass on the horizontal. And you can make this, in the computer, this two-dimensional plot. In this case, this is actually observed data. But you could also superimpose on it the calculated ones. These are slightly streakier in the reverse phase retention time, rt, just because of the scale that one uses, or the properties of the separation method.

Now, if you blow up-- if you take one of these little two-dimensional spots and zoom in on it, and look at it in greater detail, you see that it actually is a complex set of peaks in the mass direction, and fairly simple in the retention time. Now, you could say, well, maybe these are all different peptides. But in fact-- and they are, in a certain sense. But they're trivial relatives of this, where you need to think-- where you can data mine and get additional information.

So each of these are separate isotope peaks, meaning, remember, this might have 40 carbons in this peptide. And so it's going to be a binomial distribution, where this is the case where you have zero carbon-13's. In other words, it's all carbon-12's. And then the next peak over to the right is going to have one carbon-13. The next peak over is going to have two carbon-13's. And n minus 2, where n is the total number of carbons in that peptide, which would be carbon-12, the most abundant one, and so on. And you'd get every possible combination in a binomial distribution just as you would expect.

Now, what does this say? It tells you at least two things. One is the distance between these, you can see here, is a half an atomic mass unit. Well, how do we get a half of an atomic mass unit? I mean, they're supposed to be-- I mean, we know they're not perfect integers, but this is way off. And the reason is because what's actually measured is the mass over charge. You're not measuring the mass-- mass over charge.

And so this is saying this particular peptide has a +2 charge state. That's an important fact. It's going to be hard to interpret its mass if you don't know its charge, because it's m/z that's measured. The other thing that's measured is from the exact binomial distribution you can get an estimate of the number of carbons in there, because if they're a huge number of carbons, then it will turn out that one of the secondary peaks here, one of the rightward peaks, will actually be the most abundant one. If it has a small number of carbons, then the zero carbon-13 peak will be the all around winner. And so from the relative heights of 0, 1, and 2, you can estimate the number of carbons. So two facts you can get from this high resolution view of that peak.

So now we've got these two phases, these two dimensions pretty well in hand-- the reverse phase and the mass. Now we're going to add another dimension-- actually, a couple more dimensions. One is another peptide dimension, which is a Strong Cation Exchange, SCX. What this means is the peptides will have different cationic properties, different charges, and they will bind to different extents to this.

And you can put these in tandem, either physically or conceptually. You'll run one, take a bunch of fractions, and then put it on the reverse phase. The first phase literally is physically connected to the mass spectrometer. And then we'll talk about some upstream. Now, upstream separation methods, before you fragment it into peptides, you can separate. It can have dimensions on the proteins, which can tell you about the complexes.

Let's go through a specific example of complexes. So if you take the entire yeast proteome-- grind up yeast, throw it on a first separation, which is the sedimentation velocity-- remember, there's equilibrium and velocity. Velocity is mainly responsive to the size of the complex. This is a native dimension. We're not denaturing with SDS native. And so you'll see the things that go absolutely the fastest in the cell and among the most abundant are the ribosomes and the ribosomal subunits.

This is done under conditions where you just tease apart the two ribosomal subunits so that they're separated. And the bigger of the two is called 60S. The S refers to Svedberg, who was one of the pioneers of sedimentation velocity. And this can be correlated. the Rate at which it goes down, this stabilized gradient centrifugal field is related to the complex.

And then so that's the first dimension is horizontally, sedimentation. The very top axis going horizontally is sedimentation. And then going down is SDS gel electrophoresis. So now you're taking these native complexes and breaking them up into their component proteins, where high molecular weight is up, and low molecular weight is down. And you can see that there are quite a number of proteins in each of these two subunits, the 60S and the 40S.

Now, each of these proteins-- so that's the second dimension. Now, you cleave it into peptides, and you analyze the peptides by retention time in all the three dimensions-- strong cation exchange, retention time, and mass. And then you add a fourth one, which we'll develop a little later, where you can actually break up the peptides into little pieces. So you identify which protein each of the peptides came from.

Now, unfortunately, not all peptides are equal in their ionization potential. And so if it ionizes poorly, you won't detect a particular peptide from a protein. If you detect a large number of peptides from a protein, that probably means that it's abundant in your sample or your fraction, and it also means that you can believe that the computer identification of that protein is probably pretty solid. So if you get five or more peptides from a particular protein in your database search-- and we'll talk just a moment about how you do that database search-- then you believe it, and it's very solid.

And so if you look at the-- you can see that all these-- as you look at essentially the protein fingerprint for each of the 60S fractions in the sedimentation, they look pretty similar. And when you run out the mass spec, you get similar sets of peptide signatures. And they correspond to-- most of them, especially the most abundant ones, correspond to the known 60S proteins.

If you go a little bit slower, less mass, the 40S subunits and you analyze the subsequences of the mass spec signatures for those peptides, then they mainly turn up 40S. There's some exceptions in both. There are some other categories, which may be interesting. The most interesting in this particular study, the authors highlighted was [? Weimer ?] 116p, which, remember, this is a very mature field. This is a very recent study.

And ribosomes were very well characterized. And I think we had the conceit that at least in microbial systems, such as E. coli and yeast, we really understood all the proteins that were required to make a ribosome hum. But here was a new ribosomal protein, which has since been confirmed that this is a bona fide part of the 40S subunit. You can see here it had many peptide hits. So it was equal in abundance to the other 40S subunit proteins.

So here we had in a certain sense five or so dimensions, the sedimentation, which is the native complex size, the denatured protein subunit size, the peptide ionic-- the ion exchange, the peptide mass, and the peptide fragmentation. So when we talk about fragmentation, that's what MS/MS sometimes means. It means that you're doing first mass separation of the peptide. You break it up, and you do another mass separation of the component parts. And that allows us to do database searching and sequencing of the peptides. Now, how this works-- this is a blow up of something that was in an earlier slide, where you can really see the region where you've got electrospray. We'll have an even closer blow up with this in a moment.

But basically, your reverse phase liquid chromatography is going directly into this vacuum here, generating little spray at 4,000 volts, and then these molecular ions will go through the vacuum through a series of magnetic octopoles until it hits an ion trap, which helps you determine the m/z, and finally a detector. You can see there's a variety of different pressures in here, sort of increasing pressure from the point where you have the aqueous solvent going in all the way to the detector, which is the highest vacuum at around 10 to the minus fifth torr.

Now, here's where the [INAUDIBLE] mass spectrometry, or MS/MS, or collision-induced dissociation, you have-- here's your ion beam of molecular peptide ions. You've now turned peptides in-- they're each on their own little space. And in the middle of this quadrupole, this magnetic environment here, you bring in an inert gas, like argon, to collide with these rapidly moving ions, and they will break the chain, basically, at any covalent bond.

And if you think of the peptide at a chain backbone as having three different covalent bonds, there's the peptide bond itself, and then there's a carbon-nitrogen bond, and a carbonyl nitrogen bond-- sorry-- and a carbon-carbon bond. It can break at any of those three positions, and then you'll generate a set of fragments coming in from the N-terminus with three possible C-terminal fragments. And coming in, if a C-terminus is cleaving at the same point, the whole series is coming in that way.

And as you're coming in from the N-terminus, you give it A, B, and C, depending on whether it breaks at the C carbon-carbon bond, the carbon-nitrogen bond, or the nitrogen-carbon bond, A, B, and C respectively. And the same ones coming in from the C-terminus are called Z, Y, and X. And so, as it turns out, just empirically, if you sort through all the chemistry, in most cases, the peptide bond is the one that's most actively cleaved. And so the B ions and their complementary Y ions dominate the picture.

Now, the other ones will be present. Especially if they come from a very abundant peptide, they can swap out the B and Y ions for a less abundant peptide. But all other things being equal, B and Y will dominate, and most of the rest of the discussion will be about those.

This is the closest picture we'll show of the ionization step. This is a step, which has not been thoroughly enough studied, in the sense that this ionization step, where the droplets of aqueous and organic solvent coming out of the separation column is subjected to the vacuum. The water starts being released from the droplet. The protons-- remember, this acidic media associates with a molecular ion. They kind of explode because there's too much positive charge in a small space, until you finally have a molecular ion associated with one or two net positive charges. Remember, we had an example just a little while ago. We had a net positive charge of +2.

You don't need to have neutrality in this situation. Anyway, this is poorly understood in the sense that some peptides ionize much better than others. And we'll come back to this when we talk about quantification. But for right now, what we want to do is ask how do we analyze the complex spectra that comes out when you fragment-- when you take-- first, you get a fairly simple spectra, which is just a list of all the masses of all of the peptides. And remember, some will be weak, and some will be strong because of this voodoo ionization.

But then we break them. However, whatever the intensity of the original peptide was, it will make a bunch of daughter ions, which will be the B ions coming in from the N-terminus, the Y ions from the C-terminus. And you'll have this big mixture, a nested set, that get increasingly large as you get further from the N-terminus and the B ion series and their complements. And the sum of the B ion with its complementary Y ion has to be the original molecular mass corrected for the chemistry that occurs right at the cleavage.

And so here's a real example. We're going to work through it so that you can see what happens to a typical peptide here in the upper right-hand corner. And this is tandem mass spectrometry. Remember, there is a-- if you think of it, there was a single mass peptide that was then broken into all these little pieces.

And the almost intact peptide will be on the far end of the horizontal axis, which is the mass axis, close to 1,200 atomic mass units. And then a relative abundance is the vertical axis, as it's just related to the ion counts. And you can see there's some variation here. This is not due to ionization. This is due to the cleavage efficiency, cleaving at each of the bonds for the Y series, which is in blue here. And it tends to be the higher peaks, and the B series in red, which tends to be slightly lower peaks.

And then you've got these little arrows, the darker arrows indicate the Y ion series that separate two adjacent peaks. Because what's the difference between those two peaks is the addition of one amino acid. And so the focusing on the blue series, the Y ions coming in from the C-terminus, the shortest Y, the Y1, would be just the C-terminal amino acid itself, which would be arginine. And its distance from the origin would be about the mass of the arginine itself.

And then you add a glycine to it, which is a small delta, and then an alanine, and an isoleucine, and serine, and so forth. And you can see here very clearly the leucine and the asparagine and the valine, so that Y ion series all the way down. The G is the last one documented. The N and the S at the highest molecular weight are not visible. And actually, many of these things you'll have very weak peaks. You'll essentially have missing peaks that are corresponding to one of the delta amino acids. So in that case, the distance between the two prominent peaks will be two amino acids.

You can see this starting to get to be a challenging pattern recognition problem, because you've got all the B ions mixed in with the Y ions. And this is summarized in the next slide. The B ions are mixed in with the Y ions. And some of the ions are missing. Each ion has multiple isotopic forms. That wasn't so evident in the previous slide. But that blow up I showed a while back, you had that binomial distribution. There is the lingering presence of A, C, X, and Z type ions, where you've got cleavage of some of the other bonds. Ions can lose a water or an ammonia. You've got noise from other peptides and from contaminants in the system. And you've got amino acid modifications, which is not a contaminant or a bad thing. It's a good thing. This is what you're looking for. But these can be in trace amounts in this system.

Now, there are two ways to approach the awesome amount of data that you can get out of these. Remember, you've got all these multi dimensions finally ending in this forest of B ions and Y ions and all the rest of it in there. And there are two approaches. One is we'll call it de novo peptide sequencing, which would be analogous to the de novo DNA sequencing that we were doing

and the other is if you tell me the sequence then I'll find it in my data kind of game OK it's doing a database search where you're limited to proteins that are very, very-- to finding peptides that you already know about, or can hypothesize from a genome sequence. So this is the first category. This is de novo sequencing. And it takes on all the challenges in the previous slide. It takes on the possibility of missing data for a particular ion species that you think should be there, but for some reason are not efficiently cleaved by the argon in the collision-induced association.

And it takes into account that you have to have one set of B ions, a nested set of masses from the N-terminus that have to be complementary to this nested set you get coming in from C-terminus. So this is dynamic programming. And you can probably count how many different times we've done a dynamic programming algorithm in this class. And so hopefully you're happy that you did at least one of them by hand.

And this one we won't belabor, but here you can see how it kind of conceptually maps to the simplest one that we talked about at the beginning, which is comparing two amino acid sequences. There, the indels were caused by evolutionary change. Here the insertions and deletions are due to a missing ion due to inefficient cleavage in the gas phase. This is further complicated by this necessity of essentially sequencing in both from the B and the Y simultaneously and making sure that you have the best combination of B and Y assignments. So that's de novo sequencing.

Now, in slide 29, we have the alternative, which is by far more commonly applied, an example of the alternative, which is you tell me a sequence. I'll find it in my data, [INAUDIBLE]. And where you're basically calculating the spectrum that you might expect for each peptide that you might expect in a genome. So you basically use the genome to predict the protein coding regions. Use those to do in silico digestion with trypsin, basically cleave after every lysine and arginine, and maybe after some other ones are complicated rules where trypsin doesn't always cleave after lysines and arginines.

And sometimes there'll be other proteases present. And you have to take those into account. In any case, you generate a virtual set of peptides. You generate a virtual set of mass peaks. Now, since we don't know the rules that determine the height of those mass peaks, we wish we did, but we don't, so we just set it to a unit, to some arbitrary to make them all the same. So you're not going to be getting a great correlation coefficient based on the heights, but merely whether it's there or not.

And that's what you do is every time you have a hit between your predicted spectrum and the other one, no matter what the intensity of the other one is, so you waited on the observed intensity, but you have no real calculated intensity. And this correlation coefficient serves as a way of prioritizing your scores. And if you have a-- very often the best score will be the database hit for the peptide that you want.

Now, if you're expecting post-synthetic modifications, you need to tell the algorithm to add the appropriate mass to the appropriate amino acid. So for example, if you expect a phosphoserine, you have to put the phosphate mass into the program and associate it with a serine. So the serine can be either a regular serine or a phosphoserine. So that's another complexity there.

So now we've gone through the richness of the separation methods. Separation is intimately connected with getting us to a mass spectrum which is clean enough to do either de novo sequencing or database searching. Now that we've got it identified, let's try to quantitate it. We can quantitate it one of two ways, just the same as with the RNAs, either on an absolute scale or on a relative scale. What is involved?

We'll make an analogy to the RNA quantification methods, which I believe we've had something very similar to the left-hand side of slide 31 here, when we talked about RNAs, all the ways we could quantitate them. A subset of these have an analog in the protein domain on the right-hand side of this slide.

So for example, one of our favorite methods that we used was microarrays. That's the top line of the RNA. This is where you would have the gene segments, either oligonucleotides or [INAUDIBLE], immobilized on a microarray. Fluorescently label your RNA and quantitate. For proteins, the equivalent would be an antibody array aimed at unique features of each protein. This is in very early days, because we are antibody limited. We do not have antibodies against every protein surface epitope. And they are not specific enough. There's a lot of crosstalk.

We mentioned in the second line that microarrays could not measure the composition of alternative slices, or the size of the messenger RNA. So that was best done by a Northern blot, which actually measured the size, but was not high throughput. The equivalent for that for proteins is called a Western. These are all puns on Ed Southern's name. The Westerns will allow you to measure the size of native or denaturing proteins, and then detect them with antibodies. Again, antibody limited. If we had a technology breakthrough that would give us all the antibodies we need, just like we had, it's easy to dial up any nucleic acid you want just by synthesis.

There's no real equivalent to PCR for a protein world. You can tag proteins with nucleic acids and do PCR on the nucleic acids. But there's no real direct amplication on proteins. Reporter constructs basically work the same for each. You have something that is highly specific because you constructed it in vivo, but it is a sum of all the RNA and protein expression steps that give you the reporter.

Fluorescent in situ hybridization in the case of RNA or fluorescent in situ antibodies in case of proteins is a great way of correlating quasi quantitative information with a subcellular or suborganismal localization. Tag counting, there's nothing equivalent for proteins, and mass spectrometry can be used for differentiable display.

What are the sort of numbers of molecules we have? A ballpark when we're dealing with quantitation-- slide 32, it depends on the organism. Some of the simplest ones, like you find in yeast, we mentioned the messenger RNA molecules might be less than one per cell, just stochastic fluctuations. And in a human cell, it's probably a fairly good approximation or assumption that almost every nucleotide in the human genome can be transcribed. Maybe there's some leakiness where on the order of 1 in 10 to the fourth cells will have a little leakiness at any particular nucleotide. So that's kind of the background level. It's 10 to the minus fourth per cell. And it's really only achievable detection with reverse transcriptase PCR.

The entire transcriptome of a human cell is on the order of half a million transcripts messenger RNAs. And so if any particular messenger RNA got up to 10 to the fifth, it would dominate. And this happens in some cases, like reticulocytes. Maybe 90% of the messenger RNA might be globin.

Now, for proteins, you'll typically have bursts of proteins. You have one messenger RNA. You might get 10 to 1,000 proteins made, depending on the organism. And so you typically have a corresponding amplification in the last line.

Now, when people assess casually whether a particular method is quantitative or not, they can be easily intimidated. So they might say-- I've commented that the ionization of-- so ESI stands for Electrospray Ionization in mass spectrometry. If you take a protein and cleave it with trypsin, in principle, every tryptic fragment, since trypsin cleaves pretty close to completely fairly easily, every tryptic fragment, every peptide should be equimolar. If you now inject that into the HPLC, into the mass spec, every peak integrated intensity should be equal, because they're all equimolar.

And then when you find that, no, they vary over two orders of magnitude-- that is to say some are 100 times more intense peaks than others-- then you might get discouraged and say, oh, isn't a quantitative science at all. I can't deal with a factor of 100 difference. But I think that you need to reassess that when people say that mass spectrometry is not quantitative. The two requirements for quantitation is that you have reproducibility, and that you have a way of calibrating or calculating. If you can calculate from first principles, then you don't need calibration. If it's too empirical, then you need calibration.

But you do not need that every disparate object behaves exactly the same way. Not every peptide has to give the same quantitative answer simply has to be reproducible and calibrate able with that same peptide and so here's an example of establish two examples in a row of establishing the reproducibility here this is from that ribosomal protein experiment that I showed earlier with the complexes of sedimentation velocity and the multi dimensions. And this is just you do a measurement on day one, and you do the whole experiment over on day two, and then you compare the intensity of the peaks. And you get what is a fairly good straight line relationship on a log-log curve over about a little over 3 logs.

There were many moving parts in that experiment. There were all those different dimensions. And the whole experiment was not designed to be quantitative. There are no internal controls, and so forth, or so on. Nevertheless, this is a good starting point for convincing yourself or determining whether something is reproducible enough that you can make it quantitative.

Here's another way of measuring the reproducibility. That one was a correlation coefficient, a linear correlation coefficient on a log-log plot. Here is the coefficient of variation. I think we may have mentioned this before. This is just the standard deviation normalized by the mean. In the upper left-hand part of this slide, you can see that the CV, or Coefficient of Variation, is just the standard deviation divided by the mean. So you can report it in terms of percentages of variation.

So here with a calibration standards of peptides, you get somewhere between 2% and 28% coefficient of variation. That means you can trust these things to be within 2% to 28% of their absolute amounts when you calibrate them. So these two are just two examples that should reassure you that there is reproducibility and you can calibrate.

Calibration can be an expensive proposition. But there are various motivations for measuring, quantitating both proteins and nucleic acids on an absolute scale. And for example, you might want to compare them to each other. You might want to say, to what extent is it the case that the most abundant proteins result from the most abundant messenger RNAs? Or you could imagine a world where these are completely independent, since one is transcription factors. The other is translation factors. There's no reason that they necessarily are synced up. Or you could imagine a hypothesis where it's a lot of work to make a lot of protein, and so everything has to be working right for the most abundant ones. And for the least abundant ones, you can have a little more slop.

So this analysis critiqued a little bit in the subsequent paper that we'll talk about after the break can be interpreted as being consistent. When you include all the proteins, you have a very good correlation coefficient. This is a linear Pearson correlation coefficient. But as you restrict yourself to the lowest abundance proteins, it falls apart. You have less significant Pearson correlation coefficients.

So let's take a little break. And we'll talk about critiquing this a little bit, and improving, and asking other motivations for putting protein on absolute scale, and doing ratios.