ZOLTAN SZALLASI: -technologies. And this is kind of the introductory slide. You heard a lot from Zack. I think that was last time, like a week ago or two weeks ago about microarray technology. And I'm sure that he gave an extremely inspirational and enthusiastic talk about the possibilities and scope of this technology. But let me give you just a little this line.

So whenever a new technology appears or for example, microarray technology or the Genome Project, of course first, there is a general ebullience and optimism that all problems will be solved within a couple of years. There are lots of reasons for this optimism. One is that you want investors in company and you want public funding.

But of course, in a couple of months or years, realistic expectations start to appear. And then you have to start to think about limitations of the actual technology. And actually, that's the reason that we have a talk about this very topic.

So when we talk about limitations of the technology, we have to define what you actually want to do in science. Of course, there are lots of various definitions of science. But in a sense, you'd like to make predictions about some of a system. And we are going to talk about limitations in those terms. How is it going to limit your predictive power?

So when you talk about limitations, you can talk about how accurate your measurement is. There are limitations of the accuracy, of the measurements, accuracy and noise. But there are limitations in terms of sensitivity. What are you measuring. How complete is your measurement?

And of course, I'm going to be very briefly touching on this, that even if you measure everything very accurately, there are inherent limitations in your predictive power. You cannot predict everything. Think about unpredictability in terms of chaos. Even if you measure everything very precisely, there are systems that you simply cannot predict how it's going to behave in a analytic way.

So noise-- first, I would like to define what is noise and what is signal. Noise is an inherent feature of complex systems. And noise in continuous and discrete measurements, noise is the limitations of the technology. And of course, we need to talk about what can be done about noise. That's what statistics was invented for. And I'm going to talk briefly about normalization.

So what is noise? There are different definitions. This, I, of course, took from Webster. And let's just look at this point D, which is "an irrelevant or meaningless data output, occurring along with desired information." Now you should be aware of that noise is not always a bad thing. Sometimes noise may turn out to be a very important signal. And what is your background? I'm just--

AUDIENCE: Biology.

ZOLTAN SZALLASI: Biology. Sorry?

AUDIENCE: Medicine.

ZOLTAN SZALLASI: Medicine, OK-- so this is probably not the best example for you. But there were two guys, two radial astronomers, many, many years ago. I think that was almost 50 years ago, who were looking for signals. And they just saw this noise coming from every direction in the universe as radial astronomers. And that turned out to be the cosmic background radiation, which is one of the most important discoveries. And these guys that actually got the Nobel Prize a couple of years later.

That started with pure noise. And they were trying to get rid of that noise. They couldn't. And that led to discovery of the cosmic background radiation.

But if you think about medicine, then for example, the way cisplatin was discovered as a chemotherapeutic agent, what happened is actually that the electrodes that they were using in those experiments contained platinum. And they saw the effect of-- come in. Oh, you're going to hang out, OK. Actually, I just wanted to close the door because I know that's annoying.

And then they tried to figure out what was killing the cells, what was slowing down the growth. And then they realized that actually, it was the platinum containing those electrodes. And this is the way they discovered this platinum. So the point is that noise is not always a bad thing.

Now what if you see as noise or error in measurements, in biological measurements, might be a key component of biological processes. So of course, mutations in evolution are extremely important. And when we will talk about discrete measurements, one form of that is actually when you are sequencing, you'll see lots of noise in the human world, all sorts of genomes. That's called junk DNA.

Well, we do not really know what this junk DNA is for. Then-- no, just go ahead. There.

AUDIENCE: Sorry.

ZOLTAN SZALLASI: I should have said, go ahead. It's still the introduction. So we do not really pursue this junk DNA when you are trying to find genes or exons of introns or transcription factor binding site. And that is going to extremely-- it's going to bother you very much when you're trying to find these true signals in the genome. We do not really know what junk DNA is for. But there might be a good reason it's there. It might be determining the spatial distance of different genes or whatnot.

Another type of noise which seems to be very important is during differentiation. Very often, you see an asymmetric cell division. So RNA or proteins are divided or distributed between the two daughters cells asymmetrically. And that is actually done or that happens more or less by chance. And the two daughters are going to go one way or another, depending on how much RNA or protein they got.

This, you can perceive as something of a noise measurement if you do a single cell measurement. And stochastic fluctuations may be very important for the stability of complex, physical-chemical systems. I might be talking about stochastic genetic networks and robustness much later in April when we're talking about modeling. And let's suffice now, that stochasticity and noise in complex systems might be a very important feature to maintain the stability of that system.

You should be aware, of course, that genetic networks and biological systems or stochastic systems, because you know that, for example, you have only a couple of 100 copies of a given transcription factor per nucleus or even less, sometimes you have only 50 per nucleus. The intracellular environment is not a free solution. And the reaction kinetics is often slow.

And what it means is that if you have a stochastic system in this case, that if you have a completely deterministic system, then from any given gene expression found at any given state, you can go to one state, to one another state. That's a deterministic system.

Whereas in a stochastic system, from any given state, you can go to different states with a bit of certain probability. So that's what we mean by stochastic systems. Now, if you have this, you have a stochastic system in biology. Then when you are measuring gene expression levels or protein expression levels or any activity of any biological parameters, you will perceive that as a noise in your measurement.

Now is it true? Or is this really relevant to biology that you have stochasticity present in the system? And this was a paper that came out now, almost two years ago. But actually, they wanted to measure this. This was done in bacteria, but recently, similar studies were published in yeast as well.

And what they did, they took two proteins from E coli. And they put on a green, two different GFP, two different green fluorescent proteins so they could measure the expression level of two proteins. And they set up the system in a way that if it was deterministic, then it was under the very same promoter. Both genes were expressed or driven by the very same promoter.

And it was a very carefully setup experiment. So if the system was deterministic, then what they expected that the expression ratio of those two proteins is going to be the same in every single cell. Now what they found, that despite all their efforts to set up the experiment as perfect as possible, they found that two different colors. That means in this here, it's red and green. But of course, these are only false coloring. The point is that you have two different wavelengths where these signals are emitted.

Depending on the actual cell, you can have very green, very red, and some yellow cells as well, which means that despite the careful way of this system of this experiment was set up, these cells expressed the two proteins at a different ratio. And that was due to stochasticity.

So the point is that stochastic fluctuations occur in living organisms. They are trying to understand, now pretty hard, what's the relevance of this. It seems that it has a lot of relevance. But of course, we are not quite sure what the implications are.

- You showld be aware of that whenever you do a biological measurement, or what especially these days technology might develop, you're always measuring population average data. So this again, is going to add to your noise. When you are measuring gene expression levels or do proteomics, you'll of course, grind down millions of cells or tens of thousands of cells. And that is going to be giving you a certain level of noise as well.

And this is even true, this is true even if single cells are quantified. The reason for that is that if you have a stochastic network-- and let's imagine that you can really measure gene expression levels. You can do this for of course, individual proteins in single cells.

But whenever you still do your measurement, you usually interfere with the cell. Or you kill the cell. So you don't really know how would that cell have progressed. So since you interfered with the system and still have the system, you can't really figure out what would have happened in the system. So you're actually going to end up with a population average data again.

So there is no measurement without noise. As you know, it is usually the accuracy, sensitivity of your measurement. And I'm sure that most of you would be extremely troubled if you did a microarray measurement or some of a chemical or biological measurement, and in three triplicates, you would get exactly the very same number.

That would mean probably, to most of you, that there is some of a systematic error in whatever you're doing at your photometer or whatnot. Because you would expect some of a spread of your continuous data. So it is expected for continuous variables to have data with a certain spread.

And that's OK. And that's why statistics was invented. But you know that there is some of a true value of your measurement. But due to little fluctuations of whatnot, you have a certain spread around that through variable. And of course, usually the question is-- and that's what's the best statistics at this frequency statistics is really conserved with, that this is variable change due to a given treatment whenever you have a spread like that.

So if you have measurement here or here or here and this is your starting point, then what's the probability that your parameter really change its value? So what you need to do for this is need to have of course, lots of measurements and/or a fairly good idea about the nature of the noise. That's very important as well.

We are not going to get into this now. But as you know, that, for example, the easiest or the most convenient assumption is that you have a normal distribution. It's good to have that. Because if you have that, then you can actually make very simple calculations about the probability that whether your mean or whether your parameter is actually changed or not.

So statistics was invented a long time ago. And actually partly, it was due to biological measurements. And so statistics is concerned in biology with many, many different issues. One of it is what is the true value of a given parameter if there is one true value? There is a very frequently used analysis that people or biologists are not really aware of, which is kind of a Bayesian analysis. And actually, this is the most frequently done statistical analysis in biology this is the way all science works.

I have a similar belief about whether something is going to happen, or for example, an oncogene is going to transform a cell or not. I make a statement. And my job is actually, to convince you or other biologists that this is actually true. Now what you can do, what you usually do, is actually you repeat the experiment. And if you see the same phenomenon, then you are kind of updating your-- you are going to update our common belief that what I was saying that was actually true.

So Bayesian statistics is always there in a hidden way in all biology. We are trying to update each other's belief network regarding biology. The third type of statistical analysis is that you don't really believe the measurements. But you know that there is some of systematic error there. And then you try to correct for this systematic error, and that's what's called normalization.

And I'm going to talk about it in detail there. And there is a fourth issue in statistics when you actually producing a lot of, a lot of measurements. You are looking for certain patterns. Imagine that you're looking for gene expression changes that is going to cause cancer.

And you have two populations of your samples, normal and cancer. And you see that certain type of genes are always downregulated or upregulated or mutated in cancer. Now this could happen by chance. If you do not have a large number of samples and this change is actually simply random-- so in certain cells, it's up. In other cells, it's down.

Then if you have the wrong number of samples, then it might happen that just by chance, with a certain probability, in all normal samples, that mutation is not present. And in all disease samples or cancer samples, it's going to be present. So what you need to ask is that pattern, that would explain your biology be present by chance?

And this kind of, there are too many numbers. And what you can do is actually you can try to solve analytically, that's what combinatorics is about. Or you can do some of permutation. And as you'll see later, that's actually a pretty nasty problem when you're trying to apply it for real-life problems.

So biological measurements are often expensive. And something I'd like to point out to you, that if you follow the literature or when you will start to read the literature-- and I'm assuming you will because that's why you are taking this course. You going to see lots of nature science and high profile papers, in which they ran a single microarray measurement on a large number of different cancer samples. And then they are drawing all sorts of they are drawing all sorts of conclusions about which genes are important to cancer, which one is not.

And these measurements have been still rather expensive. And it's not easy to come, not easy to obtain the samples. But you should be aware of that you cannot really do any statistics on that. You should do some sort of Bayesian type of statistics. But whatever they are doing on this is not really statistics.

It's going to be Bayesian. I was going to say that I see this change very often and you either believe it or not. But you can't really get any hardcore numbers out of it that you can use for any statistical analysis or modeling. So reliable numbers cannot be produced without replicates, which is kind of obvious.

So the central problem is that in massive biological measurements, quantitative and qualitative calls are supposed to be made on a large number of heterogeneous variables, using only a few replicates. That's what you're going to see over and over again if you work on a large scale or massively parallel biology. And this is one of the problems that technology and the analysis has to overcome.

So where is the noise coming from in microarray measurements? So this is a slide, I think you saw some of variation of this in Zack's talk. So this is how an affymetrix DNA microarray chip works. So you start with tissue, and you extract RNA. And what you do is have to do an RT, reverse transcriptase treatment or step on it that is going to translate back the RNA into cDNA.

And depending on how you do it, you can either produce cDNA or CRNA. Because during this process when you are producing the cDNA or the CRNA, fluorescence dyes will be incorporated into the polymers. And then these are going to be hybridized to the specific probes present on the chip.

Now the underlying assumption or expectation is that ideally, one copy of a given RNA will produce one unit of a specific signal. If this were true, then you would have very accurate measurements. But now let's see what's actually happening in reality.

When cDNA is produced from the RNA using the RT, this is an enzyme that has its own life, its own characteristics. So the initiation of the RT step is stochastic. Because as I'm sure you're aware of that, you need a starting primer that's going to be extended by the reverse transcriptase. And very often, the reverse transcriptase, the enzyme simply drops off.

So that's why what you see is actually when you do a microarray measurement, you see usually a much stronger signal the three coming from the 3-prime end of the gene than from the 6-prime end. Because as you're transcribing and reverse transcribing the message, the RT starts to fall off. So you always have much stronger signal from where the RT is started, which is always the RT because that's where-- what you usually use is a poly(A). You're using the poly(A)-tag as your initiator.

You can use random primers as well. And actually sometimes, it's used for some in bacteria. But most of the time you start with poly(A).

Also, cRNA, which is used for the affymetrix chip, is produced in the presence of fluorescent dyes. And it's assumed that the dye incorporation. Or it was hoped or it is hoped the dye incorporation is going to be linear. And it's going to be incorporated in equal probability.

But it's not the case. The cRNA production is not linear. There are messages that are transcribed into cRNA with a much higher probability, much higher efficiency than others. And the dye incorporation is not linear either.

Also the affymetrix chip's involves a step, and you actually break down your cRNA. For whatever reason, this is the chip design. And breaking down the cRNA into small pieces is not going to be the same for all messages either.

And of course, you have all sorts of other problems like hybridization or cross-hybridization. And one can go on and on and on what would give you the noise is just a couple of samples. But the point is that your final signal is going to be the sum of all the above, or all these things and others. So this is just the give you a feel for how many individual issues will arise when you're doing a micro measurement. Of course, the surface chemistry is very important, the background subtraction, and so forth.

So let's see another example. This is the two-color microarray. The previous one with the affymetrix chip, you heard about last time, in which you are actually putting out a single cRNA per chip. And there was another competing technology invented at the same time. When you actually label cDNA of two different samples, you measure two samples, and you're actually measuring the ratio of for each individual gene.

So in this case, what you do is have equal amounts of labeled cDNA samples. And what you're hoping for, what you're trying to achieve is that if a certain message is present at the same level in both samples, then the two intensity, the signals are going to be equal. So you're going to have a kind of red, yellow spot if a gene is overexpressed or underexpressed, you're going to have a stronger red or green color.

Now what you are ending up with in these measurements is a ratio. And the problem is that actually, there is no truly blank spot. You always have some sort of a background noise there. And you are measuring the ratio. Then of course, that non-blank spot is going to give you some sort of a false pseudo signal.

So if you're for example, if there is a gene that is not present at all in a given sample and it's expressed in the other sample, then the ratio would be of course, infinitely high. Or it would be very, very, very high. But you never see this. You always have, since you have a certain background intensity, what you see is some sort of, let's say, 100-fold upregulation, which in fact, or in true, in truth might be a complete downregulation or a complete lack of that gene in one of the samples.

So this is perceived by the experimenter as compressing the signals. So you have a very wide dynamics of the ratios, from minus infinity to plus infinity, but what you actually you see-- and this is very usually, most of these measurements are, the ratios are cut-- is 100-fold up- or down-regulation on either side.

There are lots of experimental issues that can also contribute to the [INAUDIBLE] noise. So this is how the Affymetrix chip is designed. You have seen this before. And what you have, so these are very short probes-- 25 base pair probes.

The way Affymetrix tried to overcome this problem is that they designed a set of probes along a given gene using some sort of an algorithm. And what they hoped for that, if you have lots, and lots, and lots of probes-- 11 or 16 probes per gene-- then from this set of probes, you can somehow estimate the true gene expression level. So this is how they are actually designing. As you see here, this is the entire gene, and you are tiling the gene across unique enough sequence regions of the genes.

Now, the problem here is coming from the real measurement, that these up here are the perfect mesh probes that are supposed to measure the same g. To some extent, you would expect that all these expression levels, all these intensities, would be equal. And very often, for most genes, for most purposes, this is not the case. You have very bright and very dark probes.

There are lots of reasons for this. You have cRNA secondary structure, and so on, and so forth. But the point is that when you look a little bit harder, deeper, into what you are actually getting from these measurements, well, you're expected to estimate the true gene expression level from this set of intensities that can often vary by four or five orders of magnitude.

So that's how reality works, these experiments. So this is just another additional piece of information that it's not that easy to design this. But, of course, one can improve a lot, and the gentleman sitting here could tell you lots of interesting stories about how these things are designed or how they are not designed by the manufacturer. But that's the story.

So I was just trying to give you a couple of thoughts, a couple of pieces of data, about where noise is coming from in Affymetrix measurements in real life. But even if you had very good quality measurements, you have other conceptual issues in this field as well. So let's assume that you want to use your numbers to reverse engineer a system or to do forward modeling, more forward simulation, large genetic networks. But you'd like to have very good quality numbers.

The problem is that when you do these measurements, you always measure a very heterogeneous solution, a very heterogeneous population of RNA and proteins. Now, when you started this measurement, how are you going to normalize your numbers? So how do you express your measurements, even if your measurement, your technology is very good? Per unit RNA? Per microgram RNA?

The problem with this is that if you have a decrease in the level of a given gene-- and some genes are very highly expressed-- and then the message of other genes, unavoidably, is the relative increase of the level of other messages. Because let's say you have a million copies, or let's say 10 million copies for RNA per single cell. So if a highly expressed gene is downregulated, then what you perceive in your measurement is-- unless you're actually trying to normalize for the actual copy numbers-- that some of the genes are slightly upregulated. So there are other conceptual issues as well why you will have noise in your measurement.

But as I mentioned, the real problem is, the real issue is the actual technology. Now, what you can do with that is, when you have a set of measurements, you want to take a good, hard look at your data to see whether you have some sort of a systematic error in measurement. These are a bunch of Affymetrix measurements-- real life, real measurements-- in which what you see is the intensity distribution of all probe sets.

So what you have here is the measurement, gene expression measurements on about 10, 11,000 different genes, all covered by a different probe set. And this is what you see as a distribution. Now, what you see here is there is one measurement that's very strongly an outlier, and some other as well. These are pretty much the same.

And imagine that you're actually running the very same sample. Let's say you have a single cell line, and you're treating it with different drugs. Now, what you'd expect is essentially the very same distribution for each of these RNA samples with a few differences, a few variations.

And you have this guy here. So what you can assume-- and this is actually what people do and the Affymetrix algorithm does-- that for some reason during this measurement, the fluorescent dye incorporation wasn't as efficient, or your fluorescent reader was miscalibrated, or something else, but a systematic error occurred. So what you assume, then, is that the distribution for all these measures is actually the same.

So what you can do is can start to shift your curves, because you have a good reason to assume that these are actually all very similar distributions. So what you can do is actually take the mean or the median of all of these curves, and you shift them to the very same mean or median. And you simply decide where you are going to shift everything else.

And then, based on that, you re-normalize all the numbers. And when you look for differentiable expressed genes, you work with those re-normalized numbers. Because if you did not, if you hadn't done this, then you'd say that every gene is downregulated, which is obviously false. So that's what normalization is about.

So normalization, in general, is that you don't really believe the numbers that come out of your experiments, and you hope or you assume that you're going to actually improve those numbers by assuming that you have a systematic error that you can correct. There are two ways of doing this. One is that you assume that most or certain things do not change, and the second one is that actually, you have an error model.

So the first one are you assume that most are certain things that are changes to what you saw on the previous slide. So you say most of these distributions actually have to be very, very similar. And you can shape the means or medians of these curves, but sometimes the shape of the curve is going to be different as well. And, well, if you have this nonlinearity of the die incorporation, then you not only assume that-- you can assume that, if the curve is shifted, then the shape of the curve is going to be [INAUDIBLE] as well.

So you can do some [INAUDIBLE] [? lowest, ?] and you can try to change the shape of the curves as well and shape the means that most of the curves, all the curves, would look very similar. And whatever remains as an outlier after all this is done is your true outlier, what you perceived as a real outlier. And in most cases, actually, that makes a lot of sense, and it provides differential cause that can be corroborated by independent measurements.

This is the similar problem for cDNA microarray measurement. In this case, the red versus green ratios are not expected to show any intensity dependence. But in most cases, when you do two-color microarray, this is what you see. So these are the intensity, and this is the ratios. And you see that what you'd expect is a curve like this, and that is what you see.

That means that the red and the green dye is not incorporated with the very same efficiency, especially depending on the concentration or the concentration or the individual gene species. So what you see is that for let's say low-copy-number genes, red dye is incorporated with a higher efficiency than the green one. What you do in this case is actually try to straighten it out, because we are assuming that [? we have, ?] for all genes, the red and green incorporation should be the same.

So what you're trying to do is to correct for systematic errors. And in the case of when you assuming that your basic assumption is that most things do not change, then you can choose a set of elements that will be used. That is, sometimes there is a set of housekeeping genes, which is a very shaky concept. You're assuming that certain genes do not change-- let's say metabolic genes do not change or structural proteins, the genes associated with structural proteins do not change.

Now, this is used very often, as I'm sure you've seen, in [INAUDIBLE] is that they [INAUDIBLE] forget the age or [? actin. ?] Well, it's OK. It's just very difficult to find a set of genes that is really not expected to change. Or you can choose a set of special control genes that, for some reason, that those genes never change in your system.

And of course, then, the next step is you need to determine the normalization function, which is a global mean or median normalization, or some of an [? interdependency ?] normalization. If you want to learn about this more, then actually, there's a whole website, and a chatroom, and whatnot. And there's a whole cottage industry that's trying to figure out, what's the best way of normalizing a microarray?

The alternative is that if you come up with some idea about how the error is actually generated. So this is the most popular error model, in which it is assumed that at low concentrations, you have an admitted error. You have just simply a normal-- a white noise around your measurement. At high concentrations, you have a multiplicative error. And actually, for all noise, you have the combination of the two.

So if you make these assumptions, then you can generate very good error models. And the normalization based on that actually gives you a very similar result as with the previous assumption. So actually, these two methods, it seems, are interchangeable, at least for cDNA microarray.

Noise will limit the useful information content of measurements. That's the problem. That's the issue why you need to be able to deliver that. So it seems that if you take all these microarray measurements, then a reliable detection of two or four differences seems to be the practical limit.

So this is actually a very optimistic and not cross-platform comparison. So if you do a large number of Affymetrix measurements in all sorts of-- or cDNA micro measurements, or a large number of very, very different cancer samples, then it seems that if you take all the information-- or all the useful information, you extract all the useful information from your measurements-- since there are two-fold difference-- a reliable detection of two-fold difference-- is pretty much the limit, it's possible that certain genes are going to be measured reliably, with higher accuracy. But across all genes, probably, this is the experimental limitation.

And why is it an important issue? Again, getting back to the issue that you were trying to predict how your system is going to behave, let's assume that you want to figure out who is regulating whom, starting with time series measurements. So you're going to measure gene expression changes or protein changes within a certain time frame.

So how would you design your experiments? They were experiments done on the cell cycle of yeast or human fibroblasts. But, of course, you have to choose your timing correctly. So if this is the error of your measurement, this solid line, then, of course, you don't want to take measurements more often than the error of the experiment measurement allows you to do.

So if you know how fast genes are changing and what's your experimental error, from that, you can determine a sensible reliable time window, which seems to be the case that, for example, in yeast, there is no point of taking more [? geno-spacial ?] measurements more often than every 5 to 10 minutes, and in mammalian cells more often than 15 to 30 minutes. If you measure more frequently, you're just simply going to run into noise, and you're just wasting your money. So that's the reason why you need to be aware of the noise limitations. And when you know what you're error or noise of your measurement, you can make some back of the envelope calculations of how much information you can actually extract from that measurement and what that could be enough for.

So moving on to the other issue of sensitivity and completeness, when you're trying to predict what's going to happen to your system, then it's, of course, the question is that there is a trade-off or there is this issue of how many parameters are we measuring, and how many parameters should we measure? If you're trying to predict whether a certain cell is going to-- a certain cancer is going to metastasize or not, how many genes do you need for that? If you want to predict how a cell cycle is going to progress, how many genes do you need to measure for that, and how many, are we measuring?

So for that, we need to have at least some impression of how large are these networks. So this is just showing that it's pretty large. This is a graph representation of all interacting proteins in yeast. So in this case, you have about 5,000 proteins.

Proteins, genes, protein modifications are all independently regulated, so you can call them something like bio nodes. And the [? cautious ?] estimate would be that for in each cell, the number of bio nodes are going to be on the order of, let's say, a couple of hundred thousand. This is coming from the fact that you have 10,000 to 20,000 active genes per cell, and you have, let's say, less than 10 post-translation modifications per gene, per protein. And that would give you roughly this number.

Of course, that could be much more and much less in terms of whether you're working with spliced variants or you actually need to measure on the module, the activity of modules. But this is probably on this order to have such a complete picture. Now, we certainly don't have this so far, but this is the way technology has developed.

And actually, this seems to be the easiest thing to achieve. You just simply move to more and more and more genes, especially as the genome projects are being completed, and probably the coverage of the microarray chips of proteomics is going to reach a complete genome in the next couple of years. There is no real reason why it couldn't have been achieved. All you need [? to have ?] is the sequence information and set up the technology.

But so the completeness can be achieved in terms of-- if you work hard enough-- and there are tens of thousands of biochemists and biologists working on this-- you can sooner or later measure most of the biologically important parameters of the cell. At least in principle, that means that you can have a probe that would measure this. But do we actually see signals coming from these when we are using microarray measurements?

And there was a gentleman, [? Michael ?] [? Holland, ?] who did these experiments a couple of years ago, when he just simply took microarray measurements and RT-PCR measurements as well, on a couple of genes in yeast. And what he was interested in is, one thing, one, what's the dynamic range of gene expression changes in yeast? And what they found is that the [? transcriptome ?] [INAUDIBLE] in yeast carries varies over six orders of magnitude.

What, actually, this means is that there are lots of genes. There are lots of cells. The [INAUDIBLE] [? packages ?] is very large, and you cannot see this in every single cell because the lowest number means there are 0.01 copies per cell. So what you see is that certain genes, certain cells will express a single copy of a gene due to stochastic noise, and only every 100 cells will express that. So this is the dynamic range of gene expression changes.

He was also interested in that if he measures the gene expression level of these 300, 400 genes, and he chose important genes like transcription factors, and he compares the different technologies-- RT-PCR is fairly sensitive, although at very, very low concentrations you run into stochasticity, that is probably the most sensitive technology you can use. And you compare it to microarray, then how sensitive is microarray relative to RT-PCR? And that's what he saw.

And what it shows you, that this range of gene expression levels is completely-- this is not seen by microarray. So this is well under the sensitivity of microarray. What you see is that you start to see something of a correlation between the microarray measurement and the RT-PCR at two copies per cell.

So all these genes are actually expressed and changing, and probably they are doing something important. As I said, most of the genes were actually transcription factors. But they are not seen by microarray.

So sensitivity is a very important issue when you do microarray measurements. Well, then, depending on your technology, you will have lots of genes that are going to be under the sensitivity of the technology. I'm sure-- and it's, I mean, as new technologies are coming out right now, this is going to be improved as well. But this is another issue you should be aware of-- that even if you do microarray measurements and you see lots of blank spots, it doesn't necessarily mean that those genes are not changing or they are not present. Simply, your measurement is not sensitive enough.

So the utmost goal of the technology is going to be, of course, single measuring, single-copy-- sorry-- per single gene. But even if you are measuring everything accurately, there might be problems with predictions. And this is what I was referring to before.

And just very quickly, OK, because you're a biologist, so many years ago-- actually, I think it happened here at MIT-- a gentleman, Edward Lorenz, was trying to predict how the weather was going to change. It was in '60s. And what he did is he took a few ordinary differential equations, a completely deterministic system, and he tried to predict how the outcome of this set of differentiable equations is going to change.

And what he was really shocked to see-- and a little bit later, the entire scientific community was shocked to see-- that these three ordinary differential equations produced a behavior very sensitive to the initial conditions. Which means that if you just change a very little, just a smidgen of the starting parameters, the outcome of the measurements was completely different. And this ended up in scientific history as chaos theory, where you might have heard about bifurcations and so forth.

The point is that even if you start with a seemingly completely deterministic system, you might not be able to predict how that system is going to behave because of this very fact-- that small changes in the initial conditions can cause huge changes at later time points. Now, we know that biology is not like that because biology is a robust system, because we are sitting here when we are talking. So many people think that a biological system is somewhere on the edge of the completely deterministic and chaotic systems. But the bottom line is that just because you can measure everything very accurately doesn't necessarily mean that you're going to have very high prediction.

But let me give you a much simpler representation or example of the very same problem. Imagine that you already measured very accurately the gene expression level-- and at very high sensitivity-- of all genes, or many genes, in a variety of cancer samples. And what you're trying to figure out is, what are the genes that are causing cancer?

Now, let's assume that you found this subset of cancer samples that-- these are actually real measurements from melanoma. And this is-- let's say this is a subset of samples that is extremely malignant, kills the patient very quickly. And you also think that you found a group of genes that is going to be responsible for that extremely malignant state. But you need to ask the question-- as I referred to this before-- can this be due to chance, because you have a limited number of samples?

Well, just by chance, if you randomly put in those two values, you can see something like this. Sometimes you can find an analytical solution, but more often you can't. You need to do some sort of computational solution. So you permutate your data set and look for similar patterns. And if you never find a similar pattern, a similar group of genes, in the permutated gene expression matrix, then you say, well, this is not due to chance. But this is not that obvious how to do it.

So analytical solutions can be sometimes found. So let me just give you this very simple example. So this-- I usually pose a problem that you can solve at home. We had this problem that, at the dawn of microarray analysis, my lab measured gene expression measurements in different breast cancer cell lines. And when we reached-- because this was very expensive-- when we reached six breast cancer cell lines, we found that 13 consistently misregulated genes, up or down-regulated genes. And what we asked is, can this be due to chance or not?

So this was translated into a combinatorics problem that you have eight different cell lines in a gene microarray, it's the number of genes misregulated in the i-th cell line. And the question was, can we find [? K ?] consistently in misregulated genes across all these cell lines by [? chance? ?] So if you like combinatorics, this is a nice little home exercise if you want to solve.

But so you can find an analytical solution for this. And this is very simple. And this could be solved quite easily. And you a fairly reliable number of this. But what if more genes are involved? And more importantly, what if genes are not independently regulated?

The underlying assumption in combinatorics is that you're drawing your samples independently, randomly and independently. But in this case, genes are co-regulated. If a transcription factor is upregulated, well, the downstream genes are going to be upregulated as well, or some of them will be upregulated.

And this is coming from real samples. So what you see here is when you do a complete permutation, then this is going to be the distribution of correlation coefficients for each gene pair. But in real samples, this is what you see. So there is a high correlation of gene expression changes up and down-- which is kind of obvious because this is a genetically regulated network.

Now, the problem is that if you need to do this analysis, and you ask the question, is my pattern random or not, or can this be present due to chance or not, well, if you use permutated, a randomly permutated gene expression matrix as your benchmark, then in that case, your analysis, or your result, or your statistical analysis can be off by orders of magnitude-- by six or seven orders of magnitude-- relative to an analysis where you say, well, I'm going to permutate the sample but retain the overall dependency of gene expression changes. If you do that-- which is not an obvious thing to do and takes some computational tricks-- now you have a very different result.

Noise in discrete measurements. Yeah?

AUDIENCE: [INAUDIBLE]

ZOLTAN SZALLASI: OK, so what you have is that you found a pattern that a certain number of genes are causing, let's say, cancer. And what people usually do is they do a complete randomization. Let's say you just swap everybody, everybody, and then you look for the same pattern, and you don't find it. You never find the five genes that would show the same pattern, and you are happy.

Now, the problem is that this has completely destroyed-- destroyed this permutation, the codependence of genes. And in reality, that means that if you have codependence-- imagine that there are certain genes that are very strongly coregulated and other genes are never coregulated-- that the ones that are actually coregulated are not actually two independent genes, but in your analysis, that should be one, you could replace [INAUDIBLE] gene, right?

And this is what you should retain. Of course, you don't have complete coregulation and complete independence, but you have a distribution of correlation coefficients. That's what you see here.

So the way one should do this is create a large number of random matrices in which the distribution of correlation coefficients is something like this, but apart from that, it's random. And then ask the question, is my pattern present in this as well? Now, if you compare the statistical power or the statistical confidence between these two matrices, you can be off by five or six orders of magnitude. So something that is significant in this is way below significance in this.

So that's the point. It's not that obvious how to do these things. There's just an important point that sometimes, even if you have good quality measurements, biology is going to present you with very difficult problems. And this is actually present in sequence measurements. I mean the whole BLAST issue is about this as well.

So moving on to noise in discrete measurements, which is the best example. The easiest example is actually DNA sequences. So of course, you have measurement error there as well. You have sequencing errors with a certain probability. Let's say now it's probably down to 0.1%, but you use [INAUDIBLE] between 0.1% and 1%. Of course, the solution was [? sequence ?] [? a ?] [? lot. ?]

Of course, if you see a difference in your sequencing and it's not done with a single individual, you are not quite sure whether you are seeing a single nucleotide polymorphism, a SNP, or a sequencing error. But if you work hard enough and sequence enough, you will have some sort of feel about the true subsequence of a DNA sequence.

Now, you end up with a very, very, very long stretch of layers. In the case of humans, it's 3 billion. And what you need to achieve-- or what is expected from you-- is to find genes, introns, exons, the transcription factor binding sites in this sea of four letters. Now, how do you do that?

This is going to be an issue of noise as well. If you had only genes, like exons and introns, or only exons, and transcription factor binding sites, it would be very easy to find. The problem is that you have lots and lots of junk DNA or intergenic regions, and you have no idea what they are doing. And in those, sometimes, seemingly intelligible information will show up just by chance.

So how it can be found? That's why the real way of building genomes is not only DNA sequencing because from that, it's very difficult to find the number of genes. Actually, if you look hard in the literature about the number of genes, usually, the number of genes keeps falling with time because, actually, they see that there are lots of erroneous predictions. Usually, these gene prediction algorithms tend to err on the side that would give you-- on a more liberal side. It tends to give you more genes than actually is present.

So what you are looking for is actually cDNA-- for example, cDNA libraries, for the same organism, because these are the truly expressed genes. So you try to bring together the two different databases. And if you find a cDNA, well, that cDNA can help you to find the actual genes.

Now, the problem is that cDNA has to be expressed. And if you didn't happen to prepare a cDNA from the cell line in which that gene is expressed, well, then, you won't have that gene in your cDNA library. Therefore, you cannot find it in your genome.

So how it can found? And this, the DNA sequence information, can be refined to large extent by all sorts of different databases, data sources. But there are lots of unexpected issues in biology which are truly amazing. They are completely unexpected, and you would have never been able to come up with that idea simply based on primary sequence information.

And I think I'll just give you two really shocking pieces of data that are actually pretty reasonable. One is the widespread occurrence of antisense transcription in the human genome. Why do they get-- what do these guys do it, or why?

It's a long story, but what they found actually that-- they found in the human genome about 1,600 actually transcribed antisense transcription units. So you know usually how the sense-- how the genome is read and described in the sense way? Maybe just looking into whether things are transcribed in the antisense way.

I mean, you learned a lot. I mean, you learned a lot about microRNA, SiRna regulatory RNA. So there was a good reason why they were looking into this. Nobody would have expected that there is such a high number of actual antisense transcription units.

Also, when a group checked out what portion of a given chromosome is actually transcribed, they were surprised to see that it was about one order of magnitude more than expected. What people usually do is you take a chromosome-- in case they check chromosome 21 and 22-- you know where the majority of exons or introns are, and based on that, you expect that most-- well, the exons are going to be transcribed, and maybe a couple of regulatory RNAs.

So you have an expectation that, let's say, a couple of your chromosome, of a given chromosome, is transcribed. Now, what they found when they actually covered the entire chromosome by an asymmetric shape is actually 10 times as much information was transcribed from the DNA than expected based on exons. Again, you will have to predict this just simply based on primary sequence information.

But what can you do? You have this sea of information that seems to be noise. So is there a way to deal with this?

So let's assume that you need to find a transcription factor binding site. It's going to be something like T-G-G-A-C-T. Of course, you don't know that this is T-G-G-A-C-T. And, of course, it's not always T-G-G-A-C-T. It can be T-G-C-A-C-T because transcription factor binding sites like to play with sequence. And actually, this is the way they can change their affinity of their given-- or specificity of their given sequence. So then, this is going to be your actual sequence that it can bind to.

Now, so this is what you're ending up. This is what you're looking for that you don't know that this is your binding site. And you're trying to add constraints.

So this is one trick. You say transcription factor binding sites are usually within 500 base pairs upstream from the start codon of a given gene. And you also know that it tends to cluster in the same region. So for most transcription factor binding sites. You have more than one.

So what you might be doing is you say, I'm looking for certain, let's say, six base pair long sequences that tend to cluster within 500 base pairs of A-T-Gs. And then you're going to find something. But still, this is going to be very, very weak. You have way more letters, way more information, and way more noise from which, then-- the one-- than the level from which you could extract the important information.

So even if you do all this, you will find that many other of the transcription factor binding [INAUDIBLE] sequences do not function as such. Well, why? We do not quite understand yet do to higher level of DNA [? regularization, ?] whatnot. And, of course, the problem is that you do not know what sequence to start with.

So what can you do? You can hope that your statistical representation will help. And one trick is, of course, provided by nature, which is cross-species conservation.

So you have the extremely noisy-- we call them "noisy," but of course, they are not noisy-- extremely noisy genomes-- human, chimp, mouse, rat, yeast, whatnot. And you're assuming-- and this assumption is a fair assumption-- that you have a cross-species conservation of important sequences. So what you're looking for, are there sequences that are conserved across several species?

And if you combine all this with some of the smart tools like using artificial intelligence, machine learning, HMM, hidden Markov models, were extremely useful to find from [INAUDIBLE] gene identification, then you might start to see some patterns emerging.

And this was done for yeast by [? Alexander's ?] group. So this just gives you a concrete example that showed that this is actually a very efficient way to go. When they sequenced four years species-- four very closely-related sequenced yeast species-- the average number of genes in each of them were about 5,500. The reason they did it is because they knew that these were very similar species. So what they found is actually, there is a very high level of [INAUDIBLE] of genes.

So what they found is that this shows that same gene is present at the same location in all of these species. The order changes, sometimes the gene is lost, gained, especially either the chromosomes around telomers or subtelomeric regions, there's a lot of turbulence going on. But for most of the chromosomes, things are-- or the information is retained to a large extent.

Of course, there is a slow and rapid evolution. They found that for certain very important genes, there's 100% nucleotide conservation across all species. For others, there is a very low level of conservation. Probably, that's something that nature can afford to experiment with. But the bottom line is, what they were doing is actually, they found that important transcription factor binding sites are going to be present in the very same location across all species.

So what they were looking for here is-- actually, this is [INAUDIBLE] binding site-- and it shows you the four different species, and it shows this is at the very same location in all different species. This is the [INAUDIBLE], the [INAUDIBLE] binding site. It's another type of box. So it shows that you have a very high conservation of important regulatory information.

Now, what you can do is actually turn this around and look for unknown information. So what you do is-- what they did is let's generate-- or they generated all random sequences, which was X-Y-Z. That means that X, Y, and Z stands for any of A-T-C, A-T-C-G, and [? A-T-C, ?] and there is any number of random A-T-Cs between them, between 0 and 21.

You can do this. This is within the realm of scientific computation. This is actually not [INAUDIBLE]. So these are any combination.

And you look for certain statistically significant patterns for these. One of them is intergenic conservation. Are there any sequences like this, when you go through all sequences, that tend to be conserved between genes and intergenic regions? You can check for intergenic versus genic conservation, or you can check for upstream versus downstream conservation.

These are all statistical benchmarks they found for known transcription factor binding sites. So what they found, that for known transcription factor binding sites, all these have-- those are more conserved in intergenic conservation. You have a higher intergenic versus genic conservation and upstream versus downstream conservation.

So recall the problem. You're starting with any random sequence. You're just trying to figure out that any of these random sequences have any biological significance.

Now, even more importantly, what they found is that when you start to find statistically significantly retained or conserved sequences, then these motives were also arranged in front of genes that tended to share function, which is very important. Because you're assuming that there are certain functional modules, so genes that tend to do the same thing has to be turned on or off at the same time or under the same conditions.

So that's when they came up with a long list of potential transcription factor binding sites, in which all these things were pulled together. And they found that these are sequences that tend to be conserved in front of genes that tend to share function. And many of these actually were confirmed independently by experiments as new through transcription factor binding sites.

So the bottom line is that in these measurements, even in discrete measurements, this sequencing-- you will have to face a lot of noise. Biological organisms were built a long time ago, and the blueprints were lost. If you knew how it was built, then you could figure out what's important or not, but everything was experimented a long time ago.

So it seems to you now that right now, the important information is hidden in a sea of irrelevant information. And it will be very difficult-- and usually, it's impossible-- to find based on solely computational ground. But if you look for help from actual biology-- in this case, cross-species conservation-- well, then, the important things, the gold nuggets, start to emerge.

OK, and that's it. Any questions?