STEVEN A. GREENBERG: And did you go through the methods used to classify disease in that or-- here and there, OK. Well, that's what I'm focused on in the next two blocks. And I guess, this is a block of four lectures that are going to focus on this area. And this is the Use of Microarrays for Disease Classification. Oh, thank you.

And I think it's important to keep it straight in your mind when looking at papers that use microarray technology, or in your own use of it, which of these two very distinct applications you're using it for because there is some overlap between the two. But they are fundamentally distinct uses of the technology. And the two uses are for the understanding of disease pathophysiology. And that's fairly straightforward in a certain way. You look at expression of genes, and a tissue of interest, and a disease of interest, and try and understand what the mechanism of disease is in that tissue.

The other approach is simply disease classification, which can be done completely independently of understanding anything about particular genes but is essentially a strictly computational approach to looking at a large sequence of numbers that are generated by an experiment for a particular tissue. And doing this in multiple tissues and trying to classify disease, again, without reference to what the underlying biology is at all. So that's where we're focused.

And I think the title of this block is The New Histopathology. And there's a number of papers that have come out over the last couple of years that are representative. Is that going to put anyone to sleep, you think? OK.

So there's been a number of approaches to using microarray in this way and particularly in cancer. And I guess, you did hear about the approach in leukemia here. But there's a few different applications within classification.

They include things like prediction over here, trying to predict outcome, another predictive one. This one is more of just straight classification. But we're going to talk about these different methods and approaches here.

Fundamentally, I think this is similar to what happened when medical scientists started using microscopes to analyze disease. And this is a quote that microarrays have the potential to serve as microscopes to see a comprehensive dynamic molecular picture of a living cell. And so I'm going to carry that analogy through a little bit more. And we'll take a group of diseases that are muscle diseases, which is an area I work in, and look at it's classification.

And so before microscopes were invented and used in medicine, physicians did certainly know about muscle diseases and had a certain classification of them. But having the microscope and looking at muscle tissue from patients under the microscope allowed for refinements in that classification. And the way that works is you look at tissue under a microscope, and this is a cross section of muscle.

These are muscle fibers, the red one. There's some inflammatory cells here. But we can essentially make a list of the abnormal findings that we see in the muscle and use that to then classify disease.

So other cases-- yeah.

AUDIENCE: I've got a quick question So is the defining character of the inflammatory cells the fact that there are these little bodies, these little dark-staining bodies inside of them? Or are those the actual--

STEVEN A. GREENBERG: So some of these dark-staining bodies are nuclei. This is a nucleus, here. Muscle is a syncytium. So muscle fibers are tubes that have fused myoblasts that join their cytoplasm and have multiple nuclei.

And so in this particular picture here, what you're looking at, that's a nucleus. That's a nucleus within a muscle fiber. These are inflammatory cells.

They're a little bit larger. They stain a little bit darker. They have other characteristics. Yeah?

AUDIENCE: They're also much smaller too. [INAUDIBLE] So what's that surrounding the pink region?

STEVEN A. GREENBERG: This here? What this is, this is a degenerating-- a necrotic part of this muscle fiber that these inflammatory cells are invading. I can't tell you which cells they are from this view here. But they're likely to be macrophages and CD8-positive T cells that are attacking muscle.

But the reason I use this example is, we can look at a variety of features when we look at muscle under the microscope. We have-- I better turn off my auto-- we have a lot of different features that we see that are characteristic of different diseases. Inflammatory cells are characteristic of inflammatory myopathies.

Other myopathies, like muscular dystrophies, have a different pattern. There's no inflammation. There are large muscle fibers. Other diseases, like toxic myopathies have yet another characteristic feature when we look at it pathologically. And so the point I'm trying to make here is that looking under the microscope has allowed us to see different patterns within muscle and to realize that some of the things that were just called muscle diseases in the past are actually two different types of muscle disease.

So what you're seeing on the left is a sample from a patient with polymyositis, which is a type of inflammatory muscle disease. And on the right is something called inclusion body myositis, and they're mainly distinguished by the presence of vacuoles that you can see under the microscope there. And they're very different diseases. Polymyositis responds to treatment with immunosuppressive agents, and inclusion body myositis does not.

And so the microscope allows for a very meaningful classification of diseases that we were previously unaware of represented subtypes of disease. And we take that for granted now that part of accurate diagnosis for many diseases involves examination of tissue specimens under the microscope. But it's really-- the same analogy goes through with microarrays, and I just wanted to carry that through because we're not yet taking that for granted.

So in a way, the microscope lets us look at tissue and enumerate a list of features in the tissue, and they're limited. The sort of things that we can see in diseases of muscle compared to normal muscle is a small list of different things that we might see. Microarrays, similarly, allow us to take a tissue of interest and to measure the expression of genes at the level of messenger RNA for thousands of genes simultaneously and similarly provides us with a quantitative list of expression levels for all these different genes.

And so just how we-- similarly, how looking at the pattern of the microscope has helped to classify disease, looking at the pattern of numbers from a microarray experiment is also being used to classify disease. Is that clear? Yes?

AUDIENCE: You can also-- so the example you gave with the microscope, if you combine that with what we know about immunological inflammation process, it could also be used in that other branch that you mentioned at the start, the pathophysiology. So that's where the analogy is coming from, is that you can-- the microscopic classification of disease, it's more that the microscopes are the direct analogies [INAUDIBLE]?

STEVEN A. GREENBERG: Well, so what we see under the microscope also helps us to understand disease as well as to classify it. I mean, so going back to this here, this process is completely independent of understanding disease here. We don't know what vacuoles are due to an IBM.

Nobody really has much idea about why vacuoles form. Similarly with polymyositis, we know they're inflammatory cells. It's believed to be an autoimmune disorder, but we don't really have any antigens. We don't really know how it starts.

We don't if it's antigen-specific, even. But even putting that completely aside, if you just describe features, which are internal nuclei, inflammatory cells, et cetera-- just describing features allows one to distinguish meaningful categories of disease that are clinically important. There are many patients with inclusion body myositis who are being treated with steroids and get labeled as so-called refractory polymyositis before somebody pays sufficient attention to their muscle biopsy slides and realizes they don't have polymyositis.

So it's a very clinically relevant area, but it's completely independent of any biological knowledge here. And so that's why I stress this extreme. They certainly-- both microscopes and microarrays have applications in both areas. But this one area of disease classification can take place completely independently of understanding mechanism through this analogy.

So disease classification, the area of disease classification actually has a few different areas within it. And I want to formalize that a bit. And they include discovery, diagnostics, but they all have in common the concept of an expression profile representing a signature. So the expression profile is a term we use for this vector of numbers that we get from a microarray experiment on a single tissue.

So you take one muscle sample. You extract its RNA. You hybridize it to a microarray experiment overnight, and you get back a list of 10,000 or 20,000 genes and their expression levels. And we can view that list as an ordered vector, and we can compare that to other muscle samples that had a microarray experiment and also gave back a list of 20,000 ordered numbers. And we can ask questions about how similar these two ordered lists of numbers are to each other. Is that clear?

OK, we'll go through that more. So but the basic idea is that the expression-- so an expression profile is this vector of numbers. And it's the expression profile for a single tissue under a single experiment. And they're often viewed as signatures, as unique to this tissue or this to disease. And so we just want to explore that concept of signatures a little bit.

As an aside, this is the disease pathophysiology, the other branch. And I'm contrasting a bit back and forth, just to be clear. I don't if you guys have done this or had lectures about doing this and how to identify differentially expressed genes and tissues versus controls. They're fairly simple methods of just comparing ratios for doing statistical analyses. And I'm going to skip that part and get back to classic.

So staying back in classification now, the first question formalizes what is an expression profile signature and to ask whether it truly exists in a given situation. There are many papers out there that will say, we looked at an expression profile signature. But they fail to demonstrate that what they got out of a microarray experiment was a signature of that tissue other than, say, random noise.

And so there are some important aspects to a signature. And that analogy goes to that signature we use when we write. And signatures that we use are useful because they're distinct.

My signature is different than yours because it's reproducible. Whenever I write my name down and my signature, it's pretty similar in that it's readable in a certain way. That's not always the case with signatures.

This is similarly true of what we would call classes. And classes are groups of conditions that fit into one category of some sort. In this context, in the context of using microarrays to classify disease, they are, say, the expression pattern-- the expression profile from 10 patients with a single disease. Does that form a class, those 10 expression profiles together? And classes need to have similar properties as signatures.

OK, I don't know why I put that up there for that one. OK, so this signature idea, again, is, that's my signature, and that expression profile would represent the profile of a tissue specimen under microarray experiments. And so the first question is, is it reproducible? And so when you do look at papers in this area and you're interested in data, you need to see that somebody can do the same thing twice on a given tissue and get the same pattern out.

I could take bucket of paint and throw it on the wall and call that a signature, but it only is meaningful if I can reproduce that and get the same pattern when I throw the bucket of paint on. So you do need to show that you can, and, certainly, it is possible to do it. This is data from our experiments. But you should see this type of data when you look at papers that people have reproduced the profile they get out, done two experiments on their tissue sample and gotten the same pattern of numbers out each time.

It needs to be clear that this is distinct, that the signature from, say, one disease is different from another. And there are ways to do that. And then the question of readability comes out. And that's, how do we read a signature?

So if I do an experiment, I have these vectors of numbers. And I'm going to do this for 10 different tissues and compare and ask, how are these vectors similar to each other? We have to define some way of comparing vectors to each other and judging their similarity. These are 20,000-dimensional vectors here.

And there's a number of different measures that are used to look at similarity of these profiles. And they include use of correlation coefficient-- so just taking the Pearson correlation coefficient of these two vectors. And that's what I did here.

This is just the Pearson correlation coefficient of one set of microarray numbers against another. Euclidean distance-- mutual information has just about dropped off the scene at this point as a use in this field. But there are a number of similarity measures, and this will change how one reads these profiles and compares them.

And so the same principles that apply to signatures also apply to classes. If we define classes, the classes also have to have the same features. And the way we look at distinctness of classes are mostly through clustering methods. So clustering methods, which you may have had some exposure to here, what clustering methods are a way of organizing tissue samples by similarity into classes.

And they differ substantially in terms of the way that they create a organizational class structure or classification. So hierarchical classification is used quite a lot, and these are trees, basically, of relationships. They're not actually true classes in the sense that they're distinct. Everything is related to everything else under hierarchical clustering. But the degree of relationship is just greater for profiles that get classified together on the same tree, the same node.

There's different types of hierarchical classes. There's non-hierarchical classification. And so the point here is to emphasize there's a variety of mathematical methods that are still used in this field to try and define distinctness of classes. Reproducibility-- so again, the signature issue was distinctness, reproducibility, and readability. Reproducibility of classes is a big issue here, that when you see papers which will look to say we have a certain classification.

And like the leukemia one, we have ALL and AML, separable into two classes. It's important to ask whether this is a reproducible classification system, that if you do this under some slightly perturbed way, whether you end up with the same classification. And this is a somewhat neglected area still within this field of demonstrating reproducibility of classes. Readability-- we mentioned the different similarity measures that are used to read out expression profiles. This is just the correlation coefficient.

OK, so we're going to go into a little bit more detail about a hierarchical clustering because it's such a commonly used technique now and how tissues are organized by similarity. And the basic method for this is to-- so again, our basic data set consists of, let's say, 20 different tissue samples and the expression profile for each one-- so 10,000 genes measured for each of 20 different profiles. And we're going to try and organize these 20 different tissue samples into groups or some structure based on their similarity of their expression profiles.

And so the way hierarchical clustering works is to first look at all pairwise correlation coefficients of these 20 samples. So sample number 1 and sample number 4 have a certain correlation with each other. And 1 and 5 have a certain correlation with them. And we're just going to calculate all correlation coefficients. That's the first step.

And then for the single highest correlation coefficient-- so the two samples that are most correlated, it might be sample 8 and sample 15-- we're going to define those as being together, as close together and being two leaves, basically, on the end of a branch of a tree. And then we're going to repeat this for the remaining samples and build a tree up, basically, from the leaves. This is one type of hierarchical clustering.

So here, there's scatterplot diagrams for three of the pairs within a data set that are shown. But you do this for all pairs here. And then you start-- let's say, sample 1 and 2 were the most correlated from the whole sample.

You join them together as leaves. And then, let's say, the next ones with the highest correlation coefficient are sample 7 and 5. They get joined together.

And then the next highest correlation coefficient might be the average coefficient of the S1/S2 group and the 75, so they get joined together. So this tree gets built up. And I guess, I'm assuming you've seen some of these pictures of hierarchical clustering diagrams. Have you seen any pictures? OK, well, maybe I'll show you one.

Another way to look at this process is to switch to colors. And this is a pretty common visual representation in this field of how to represent these data sets. So what I have here is, my columns are each a different tissue specimen. So they're different, let's say, muscle biopsy specimens from patients. And for each one, we did a microarray experiment.

So there were seven microarray experiments done. And we measured 12,000 genes and their expression here. And we put this into a table and then color code it according to some color scale so that the greens represent the lowest expression levels and the reds, the highest. And so we end up in this color diagram here.

And so the process of them classifying our tissues is a process whereby we try and find specimens that have similar colors at each gene. And so classification involves shuffling these columns. One can also classify rows.

But for the purpose of classifying disease, we'll shuffle the columns accordingly. And when you do that using hierarchical clustering or other algorithms to do that, this contrived example, there were three different classes represented in here. And that's what disease classification is using microarrays.

One can also shuffle the rows and classify the gene or cluster the genes here. What that tends to do is to bring genes together of similar function within the rows so that genes that are behaving similarly across this data set of tissues will tend to get grouped together. And that's a different-- there's things you can do with that. You can infer gene function and other things. But it's not what we're focused on. Yes?

AUDIENCE: I got a quick question. So this is just a representation of a microarray experiment?

STEVEN A. GREENBERG: Of a group, a data set made up of seven microarray experiments.

AUDIENCE: Right. And when you're choosing your tissue specimens, the seven that you're looking at, are those the actual, I guess, specimens from tissues that you suspect for now has a particular clinical phenotype of a disease? Or do you throw in your control tissue specimens within this, basically, to demonstrate that there isn't?

STEVEN A. GREENBERG: It depends on what you want to study, what you want to demonstrate. And so there certainly are-- in the diseases I study, there are patients whose muscle biopsy looks normal under the microscope, even though they have a muscle disease. And so that's of interest to us then to do a classification experiment here, where we take that data from, say, five patients with a disease, but they have normal muscle biopsies, and five patients with normal muscle biopsies who don't have a disease, and to shuffle them, and see if they fall into two classes properly.

And they do. I'll show you an example of that. So in that way, we can demonstrate that, sometimes, microarrays can see things that we don't see under the microscope. But it strictly depends on a particular application. The principles are what we're focused on here.

AUDIENCE: And when you're doing these samples, are you doing them more in the sense that you're organizing these according to either row or column before you know what's going on?

STEVEN A. GREENBERG: Yeah, they're being done strictly through, say, an algorithm like this, hierarchical clustering, where you're just looking at the columns of numbers, and you're saying, this vector is similar to this vector. Let's put them together. This other vector is the next similar most. Let's put that together. So it's completely independent of the labels.

So a lot of people think of this field as supervised or-- so this particular application is unsupervised classification. Have those terms come up, supervised versus un? It's the way a lot of the artificial intelligence and machine learning community thinks about this area. I don't like to think about this area in that way because that's secondary whether a method is unsupervised or supervised.

That hasn't helped. That's confused me so many times that I like to think about it in the way that I'm presenting it to you. So this is disease classification.

So there is a group of distinct methods that we call class methods that are applicable for a number of different types of applications within disease classification. And the three methods are class comparison, class prediction, and class discovery. And so class comparison is an area in which we are simply describing the different classes that might exist within one of these data sets that we put together from-- let's take, for an example, 50 samples and a microarray experiment for each of those, and consider that one data set.

And so class comparison is a exploratory method to just compare and contrast whether there are classes within this data set, what the differences are. There's a variety of computational tools that are used, and these include cluster analysis, the so-called supervised learning, fold analysis. But I'm focused more on not the computational tool that's being used, whether we're doing unsupervised or supervised classification, but the broader concept about what we're trying to do.

And so one group of applications, we're just trying to explore the data, and its class structure. And that's called comparison. Prediction is a whole different area. And prediction is a method by which we want to predict the class of a new sample that's given to us based on our prior data.

So that involves using data that we have and constructing some function, essentially, or some method based on that data set so that if we encounter a new sample, that we can then predict something about that new sample based on the prior data that we had. That's prediction-- very different than class comparison, although some of them use similar methods. So the tools that might be used to do class prediction can also be used to do class comparison, some of them.

Class discovery is yet a third application. And class discovery is focused on discovering new disease subtypes. So it's most analogous to what I started to talk with in terms of looking at things we learned when we looked under the microscope, initially, about muscle disease. We learned that there are different types of diseases that truly are subtypes of muscle disease that we didn't know about beforehand. And so class discovery is focused on looking at microarray data from tissue samples and trying to understand if there might be some subtype that's defined by the microarray data that we weren't aware of beforehand.

Such a subtype has to have some meaning to it. It can't just be that we found, looking at these very undetermined data sets, that we found different classes of these vectors. It has to translate to something meaningful. So one has to define classes and then go back and say whether this class actually means something, whether these patients who fell into this class are really different than the other patients, and people just never noticed that before, all right?

So going through those, class comparison, again, is used to establish distinctness of classes and just compare and contrast. There's no direct medical applications of this tool. There are indirect applications. It can be useful for generating ideas about classification and pathophysiology. And there's a variety of computational tools that can be used to compare classes here.

This is an example of class comparison where we looked at 45 muscle biopsy tissue specimens from four different groups. We had normals. We had patients with Duchenne muscular dystrophy, nemaline myopathies, and then a group of inflammatory myopathies, of which there's a variety of subtypes. And one of the tools for class comparison is hierarchical clustering, which we did.

And when we used hierarchical clustering applied to this data set using one of these color schemes, this is the raw data here in this column. And again, this is 45 columns here. And this is just a portion. Maybe this is 3,000 or 4,000 genes that are represented in rows here. And you just look at it, and you can see streaks of green or red going down, suggesting that there are certain classes here, that certain tissues seem to have, say, low expression levels for these genes and high expression levels for these genes.

This, here, is a blow-up diagram of this particular section here. And it has the labels. And so the hierarchical tree here has four branches, essentially, here. This branch is all of the normal specimens. So all the normal specimens got classified correctly here.

These are all the inflammatory myopathies. These are mostly patients from a third group, nemaline. And these are mostly patients with Duchenne muscular dystrophy.

But this is class comparison. We just applied hierarchical clustering to this data set and found classes here that we're interested in just looking at. We knew about these classes already. Yeah?

AUDIENCE: The reason that this is not a direct point of application is because you're not going in and trying to-- I'm trying to make connections between these different patterns that have allowed you to classify these or make these different--

STEVEN A. GREENBERG: Well, it has indirect applications. But by direct, I mean, very direct. So that if I have a new patient, and I do a microarray experiment on it, can I make any predictions of that based on this approach that I've used so far? Or have I discovered any new classes of disease here amongst these groups? Those would be fairly direct applications.

There are certainly indirect applications here. So there are things that come up. Like, the only misclassification for the inflammatory myopathies was this one, one patient with an inflammatory myopathy. This one, who had this disease called IBM, got classified in this group instead of this group. And it turns out that this one was really different than all of the others and that it was a patient who had a familial form of IBM.

We don't know what genes are involved, if any. And we don't know what this is. It's very rare for familial instances of IBM.

But this particular patient had familial IBM. So this might tell me that there really is something different about that disease because under the microscope, we don't see any difference in that patient compared to the other patients with inflammatory myopathies. So it helps you generate hypotheses and ideas.

AUDIENCE: I guess, I'm just confused because, in this particular example that you've given us, it started as something that you knew [INAUDIBLE] your data set. So there was no new identification of the disease--

STEVEN A. GREENBERG: Correct, right.

AUDIENCE: But given a sample of tissue samples that were from patients that had diseases that affected something else, you didn't know about this [INAUDIBLE], it seems like could have this classification of new disease.

STEVEN A. GREENBERG: You can use this tool to do that, but you have to add in a few other things. Yeah, so I mean, this tool of hierarchical clustering does cut across the three different major class methods-- exploration, prediction, and discovery. But you have to add more to it.

So I didn't mean to confuse you and say that way. But you do want to keep the tools separate from the method, the overall approach that you're doing. If you're just exploring data, there's a bunch of tools. If you're trying to make predictions, diagnostic or prognostic predictions of new patients, it's an overlapping set of tools but a totally different approach or a different procedure.

So yeah, let's skip this one. And that one too. OK, so class discovery, then, is the next area of those three methods that I outlined. And this one uses expression data to discover previously unrecognized but clinically relevant disease subtypes. It also doesn't have direct medical applications.

And by direct, I mean things that would immediately be applicable to patients-- and I mean, patients who you might see in the office. And maybe I should revise that one here. I haven't over the last year, but maybe. But certainly, indirect, it does advance the field because it improves our ability to recognize diagnostic and potentially prognostic subtypes and potentially treat patients differently based on this subtype of disease that they may have that we didn't previously know about.

And so the basic method of class discovery often starts with cluster analysis to try and define classes. And it can be any hierarchical or K means-- any variety of approaches to defining classes here. But after that, one explores the phenotypic variations within the defined classes for a number of different types of phenotypic variables.

So you take 50 patients with, let's say, a type of cancer, who you think all have the same type of cancer. Our best diagnostic classification system now of looking at history, and physical exam, and doing tests says that all these patients are indistinguishable. And you do microarray experiments, classify the data, and then you'll will certainly get out classes, whether they're real or not or meaningful, doesn't matter.

But the algorithms are guaranteed to spit out classes for you. And if you then want to go back and ask, well, is there something different about this class compared to that class that's clinically meaningful? And one thing might be time to some endpoint, such as survival.

Do patients in one class, based on their expression profiles, live longer than patients in a different class? Or any other endpoint-- do they not respond to treatment? So I go back and I see there were 20 patients in this class and 10 in another class and say, how did the patients in this class respond to treatment compared to the patients in this class?

And this is actually a fairly powerful method to discover new disease subtypes that we don't know about. Yeah, I think it's an excellent method to use. It certainly has been successful in a variety of areas and cancer, for sure. Its application to other diseases is lagging behind that of cancer.

There's a few examples where it's been done. I think I was planning on going through a couple. I don't if we need to. Do want me to go through an example? This is a lymphoma paper as an example. OK.

In this paper, the investigators had 96 patients who they knew about nine classes ahead of time. And these are all patients who have different types of lymphoma-- 1, 2, 3, 4, 5, 6, 7, 8, 9. Actually, there's eight lymphomas in one normal class here.

And they used lymph node tissue from patients who have lymphoma from some normals. They did microarray experiments. So they did 96 microarray experiments, measured something like 4,000 or so genes for each of 96 patients, and did clustering of the data, hierarchical clustering, and then for purposes of presentation, colored things in for us here.

So this is the dendrogram structure that they obtained. And they've nicely colored in the picture so that all patients who had one type of lymphoma, CLL, are yellow. That's this group of patients.

And the patients with this type of lymphoma, diffuse large B-cell lymphoma, are coded in purples. I think one or two-- one of them ended up out here. One is here. One is here. Most of them are in this section of the tree with a couple of other diseases mixed in.

So at this point, what would you call this at this point of the class methods? Exploration, discovery, or prediction? Comparison, right. So this is class comparison they're just looking at here. But they go on, and they focus on this structure here for this disease. So looking at the patients with diffuse large B-cell lymphoma, they noticed two different branches at this level. So there's this group of patients starting from this one here out to here. And then there's another group starting from here out to here, OK?

And so then they did class discovery. So they asked, that's curious, why do I have two different branches here? And it may just be an artifact of the algorithm. In fact, again, hierarchical clustering doesn't really give you classes.

Everything is continually divided into a binary tree structure. And so it's always going to split things down and you have to decide what level you want to look at similarity. But they ask the question. So they then went and said, well, what about survival of these two groups? Do the patients in this group here have a different survival than patients in this group?

And so let's make sure that I have that. So what they show here is that the diffuse-- so I've lost the indicator. Hang on one second.

Let me make sure that I know what I'm talking about. Yeah, that's it. Yeah, right.

So they found, if you looked at survival based-- so let's ignore this one on the left and look at these two here. So survival and decisions about chemotherapy are based on largely on something called the International prognostic index in this disease. And that's a measure of clinical risk.

And they found that-- actually, let's do this one first. So they did look at survival for the two groups and found that they did have very different survival here, that, one group had much better survival expectancy than the other group. They then further looked at the International prognostic index and how it grouped patients.

And according to that scheme, there is a group of low and high clinical risk patients. And the high ones will generally get more aggressive treatment than the low ones. And so they looked at this low-risk group according to their microarray patterns and asked, how did they fall into these two different groups that we found? And they found that a number of those patients-- 14 of them-- were in the Better prognostic group of the two they had defined. And another 10 were in the worst.

And so they were able to refine and say that this group that's previously been called low clinical risk based on the international prognostic index actually has two subgroups within it. And one subgroup has a pretty good prognosis, and the other has quite a bit more aggressive disease. And they averaged out to this, right here. But the microarray data suggests that these patients are at greater risk and should probably be treated more aggressively based on their microarray data generate expression profile.

AUDIENCE: [INAUDIBLE]

STEVEN A. GREENBERG: That takes age, and gender, and stage of disease at presentation, whether it's just in the lymph node, whether it's metastatic. There's a few other variables in there. And I'm not an oncologist, so I'm not sure.

But there are clinical variables. There are patterns under the microscope as well. But it's the best that medicine has.

And the microarray work suggests that we could do better. And that's not an unreasonable thing to suspect in the first place, particularly in cancer, where, if we can generate 20,000 numbers that represent physiology-- and that's the other thing is, the microscope really doesn't show physiology. It shows anatomy there. The microarray is-- and that's what the quote said at the beginning-- a physiological microscope looking at living processes. Yes?

AUDIENCE: I've got a question. I get the impression that there are a couple of marks there. When I look at the y-axis and I see probability, could you comment on what that measure actually is?

STEVEN A. GREENBERG: That's the probability of being alive at a given year, at a given period. So at four years on the x-axis, you have a 20% probability. Oh, no-- and this curve being alive, yeah. These are actually censored data here.

And that's a complex thing about data representation. I didn't want to get into that, except to the point to take out of this, I think, is that this has enormous potential for disease discovery of subtypes, disease subtypes. And again, particularly in cancer, where the physiology of these tumors, what genes are turned on is probably very much related to the disease course. Do these have metastatic potential? Are they growing rapidly or slowly based on the expression of genes in the tumor?

Yeah, that was the point about adding to the IPI. OK, I won't do another example of class discovery. There's plenty out there right now in a variety of different areas. OK.

And there's class prediction. So class prediction is a method of using expression data to build a model that will then predict class assignment of a new sample presented to that model. This one has potential very direct medical applications. If one can build such a model for breast cancer, then when you see a new patient with breast cancer and you do a microarray experiment on the cancer tissue, and apply it to the model, and the model spits out some prediction-- high risk of death within the next year. That will help you to decide how to treat the patient, if it's accurate.

This can be used to establish a diagnosis, to make predictions of outcome is one of them. But it could be things like predicting response to a medication. This patient is more likely to respond to this medication as opposed to that one, based on their expression profile. Indirectly, it does tell us about disease pathophysiology, to some extent.

So this method is a bit more involved, but it's pretty standard in the way people do this now. And so the approach is to take a data set that's sufficiently large-- let's say, 100 patients with breast cancer-- microarray experiments on each of the 100 cancer specimens for 20,000 genes, and that's the data set. So the first approach that's usually taken is from, within that data, to choose a gene set that will discriminate amongst classes.

So in this approach, one decides ahead of time what one hopes to predict. So let's say, one wants to predict good outcome. A good outcome might be survival five years from now without metastases-- does a disease-free survival in five years from now.

And that's a good outcome. If patients have that, we call that good. If they don't, that's bad.

And so we're going to try and build a model based on our data set to predict good versus bad outcome. So this will be a binary predictor. We're trying to predict one outcome or the other. One can build predictors that are not binary that put people into one of three classes or four classes and say, you know, so forth. But this is the simplest type of binary prediction.

So instead of using 20,000 genes, we're going to whittle them down to a more representative set of genes that are more meaningful to build this model, and that's called the gene set. Next, we have to construct or choose a prediction function that when applied to a new expression profile will produce a real number. So this prediction function is going to be a mathematical function. Take this times that, add this, square it, subtract something else. And it's going to spit out a single number-- 8, 8.5, 6.2.

And then we need to choose a prediction rule that will classify a sample based on the output of the prediction function after application. So we take our new sample, our new vector, we feed it to the function. We get at 8.5, and we decide if the number is greater than 5, it's good. If it's less than 5, it's bad.

And then the last method is to validate this model and its application. And that's class prediction. So going through it in a little bit of detail, we start with our data set on the left.

We have columns. So we have 20,000 rows, whatever number of rows here-- I guess, 7,000-- some number of rows here. And we have our columns, which represent individual expression profiles from different tissue samples in patients with breast cancer.

We figure out some discriminating set-- and there's a variety of methods to do that. We'll go through a couple. And we decide, well, we're not going to use all 20,000 genes but maybe just 500 of them.

And we pick those genes. And then we construct our prediction function. And the prediction function looks like that.

So if I feed it a vector, we're going to take the expression level for gene 8 and multiply by 2 and the expression level for gene 33, and square it, and do all these other things. And the prediction function is going to give us a number, like 8.5. And then we need a rule.

And one type of rule is a simple threshold rule. If it's less than 10, it's in one class. Greater than 10, it's another class. All right? Make sense?

Now, there's plenty of options for building these models along each step. And there are papers using all of these different options to build these models. It's still rather ad-hoc approach to model building. One approach is to cluster the genes-- so to do class exploration-- cluster the genes and the columns.

And look at it and say, there's a lot of red for these genes in one class and a lot of green over here. I think this set of genes are important. I'm going to use that in my discriminating gene set. That's a pretty successful way of doing it, actually.

Another method is something called correlation to ideal outcome. There are other principled methods for doing this, for finding the discriminating gene set. This is one where, from my data set, if I was trying to build some of predictive model, this is the clustering of both the muscle samples and the genes going that way.

And I'm looking at IBM and polymyositis, and I see a lot of red and a whole cluster of genes that are overexpressed in those diseases compared to others. The genes have a lot of similar function. They're immunoglobulin-related genes, mostly.

And so if I was trying to build a model that would predict a new sample as either being within this group versus something else, I'd probably use these genes in the discriminating gene set. That would be one way to do it. Same thing here, and one can do that.

Another option is what's called correlation to ideal outcome. And this one is used fairly often and seems like a pretty good method. And the way this works is, we have our samples, and we have our genes. And we will first organize our samples into the classes that we want them to be in, the good versus the bad. So take all the breast cancer patients who had a good outcome at five years, and we'll put them in one group, and the ones that had a bad outcome in the other.

Then we'll create this ideal vector in which ones represent a good class, and zero, the bad class. And we'll look at pairwise correlation coefficients for all of our genes compared to this ideal vector. So we essentially-- correlation coefficients test linear direction of correlation. And so we'll be asking, which genes behave like this vector? Which genes are up in this class and then change and down in these other class? And that's how we'll find our gene set.

AUDIENCE: [INAUDIBLE] question of 1 or negative 1 [INAUDIBLE]?

STEVEN A. GREENBERG: It doesn't matter what you use here. You just use two different numbers. And you can do the math. I mean, linear correlation coefficients are invariant to stretch transformations. So I don't think 1 or 0-- well, 1 or--

AUDIENCE: [INAUDIBLE] in the other case. So if you separately tiering down, you could also [INAUDIBLE] exactly how it [INAUDIBLE].

STEVEN A. GREENBERG: Right. You can use any ideal vector as long as it's one numbers represent one class, and a different number represent the others. Yeah?

AUDIENCE: In the beginning of the lecture, you mentioned the binary classification method. Is there any way to expand it to fuzzy logic or multiples? In your own sample, it tests several of the test [INAUDIBLE]

STEVEN A. GREENBERG: Yes. You certainly-- it just matters how you build the predictor function. So at this step for discriminating gene set, this could certainly be extended. This particular method of finding an optimal gene set, correlation to ideal outcome, could be established for any number of groupings of classes. You could use 2, 1, 0, different numbers in this ideal vector.

And you will end up with a discriminating gene set, a set of genes that do have some differences in expression across the three classes. But I don't know. I'm sure that people have applied fuzzy logic.

People have applied just about every mathematical method available to doing classification and model building. But I don't know. If that's something you're interested in, you might want to look at that and see what people have done.

So but this basic method is to correlate all of our genes to this ideal vector to compute for each one the correlation coefficient with our ideal vector. And I listed fictitious examples there. And we see that, say, for gene 2, it has a very high correlation coefficient with the vector. So we might want that to be part of our gene-discriminating set to build this model. So we might take the highest 100 genes out of 10,000 here, the ones with the highest correlation coefficients, and use that as our discriminating gene set to build this model. Let's say, these here, OK?

OK, so then we want to make a prediction function once we have a discriminating gene set. And again, lots of ways to do this. How to compute a function that will give us back a number that differs significantly between the two classes, like that function.

And then lastly, a prediction rule. The rule is often done as a threshold, saying, like that example there. If it's greater than or less than one number, it's in a different class.

There are certainly prediction rules that can be ambiguous as well as deterministic ones. So you can have a prediction rule that says, if it's greater than 10, it's in one class. If it's less than 5, it's in another class. And between 5 and 10, I can't decide.

And the choice of the rule is the classic trade-off of sensitivity versus specificity, that once you build your model and your prediction function, you can then arbitrarily vary your rule to optimize sensitivity and specificity. Depending on, let's say, a threshold, depending on where you put this threshold, you might get all of them correct in one class, but there are some from that class that were in the other class incorrectly. Or you might move it down some other way, and then you capture everybody who belongs in class 1, but there's a few people from class 2 who are also being classified in class 1. Everyone OK with that?

And that's an example here, without going into too many details, of classification, that the black is one type of sample, the white is another. They built a model. They used a discriminator gene set, the function.

And then depending on whether they put their threshold here or here, if you put it down here, you get more of the blacks into this classification, but you pick up a couple of extra whites here, which don't belong. The blacks belong up here, and the whites belong here. And so depending on where you move this line, up here, you get different values for sensitivity and specificity of your predictive model.

OK. Now, the last part of this for class prediction is validation. And this is a very important part. So the question is, why do validation? And that's my only fun slide there. And that's because of overfitting.

So the problem is, let's say, you have a data set here. And I want to build a prediction function which will tell me what to do with new data. So I could draw a straight line through it. And now it's-- I'm going to use a simple model like that.

Now if you give me an x value here, I'm going to predict the y value using the function, and that's right there. Now, let's say, instead, I start with this data set, and I use a very complicated model instead of a straight line-- one that works perfectly, in fact. Every data point ends up on the 100% correctly predicted by this function. But this is some complex function. This is more than a cubic here, but that's what I put up as an example.

So if you get a new point now and the x-coordinate and you ask for a prediction of y, you might get something down here from this function, which probably is not the best match. There's no one really to say what the best match is here, but it's probably not the best match. And this is the issue that continues to plague model building in this field.

And that's the overfitting because what these models are doing are taking 100 samples, let's say, 10,000 genes for each, and picking out of these 10,000 genes a set of genes and a predictive function that correctly classifies this set of 100 samples. So you essentially have 10,000 or 20,000 variables to use to define 100 samples into two classes accurately. And it's a highly underdetermined data set. If you allow yourself the freedom to choose any one over number of ways to pick the discriminator gene set to build the function, one can take random noise and build a perfect predictive function that will classify it.

And so that's where validation comes in. Validation is an approach in which one has built the model and then needs to test it before deciding, I'm going to use this in patients to decide how to correctly predict new samples. And the correct way to test it is to, once you have your model, is to then look at your next 100 patients who come in, and do their gene expressions, make a prediction, wait five years, and see who is really in the good versus the bad outcome. And that's the way this will need to be done.

But in the meantime, nobody really has time to wait five years or to do this on another 100 samples of patients and to look that far ahead. Although I suspect people are starting to take that perspective in this.

So what people have usually done is they go back to the original data set they used to build the model and validate the model on that data set, sort of. Sometimes, this is done through the combination of a training set and a validation set. So you had 100 samples.

So why don't we just build the model on 66 of the samples and save the other 33 to test it? And that's fine to do that. So people build the model based on two thirds of the data and call that the training set. And then the other third is used to validate the model and test its accuracy.

Another approach when people don't have a separate validation set and use all specimens to create their model is called cross-validation. And that's a way of validating the model on the original data set that was used to construct it. And there is a potential way to do this OK. It's inherently got problems with it.

The most common method is leave-one-out cross-validation. So in that method, you have 100 samples. First, you remove one sample. So now you have 99 samples. You build the model based on those 99 samples, and then you test it on this one sample that's left.

And you see whether it makes the correct prediction. Then you do it again. You take another sample and remove it, build another model based on those 99 samples, test its prediction for that sample. And you do that 100 times, and you ask, what is the accuracy of this approach? And did it accurately predict all 100 times correctly the class or not for the sample?

So that's validation options. So that's the leave-one-out cross-validation construct predictor. Apply predictor to one left-out sample, repeat for each sample, calculate an error rate as present misclassifications.

So the way that cross-validation works is, again, the underline there is construct predictor using all data except one. If we go back to the method we use to construct a predictor is, from the data, we choose a gene set, we choose a prediction function, and we choose a prediction rule, right?

So a common problem in this area is that people don't do this. They skip step one. So they don't go back and rechoose a new gene set each time they apply cross-validation.

If you think about it, you will have to choose a new gene set, and you might get a different set of genes each time. And that's a problem. If you're trying to build a model and say that these are the genes that are important in my model, I'm going to use them for new patients, you can't really test that hypothesis using this approach.

And so people have looked at simulations of doing incorrect cross-validation of not repeating that first step. And essentially, the simulations show that you can build predictors that are 100% accurate, even in a data set that's completely random. All right, now, I want to wrap this up.

But if you repeat the process correctly by reselecting the informative gene set each time you leave one out, then the predictor's accuracy is no better than chance, as expected in a random data set. So it makes a huge difference here.

So many papers have incorrect cross-validation in them. That paper on breast cancer and nature and its follow-up in *The New England Journal of Medicine* as well reported an accuracy of 73% of its predicted. They built a predictor to predict good versus bad outcome and reported a 73% accuracy.

This is based on leave-one-out cross-validation, but they did not recalculate the informative gene set. If one does that, the corrected actually is only 59%. It's not much better than flipping a coin.

Other papers, even the ALL/AML did incorrect cross-validation. Well, that's not correct. They didn't do incorrect cross-validation, but they had a classifier which was ambiguous.

That's another issue in reporting accuracy, is they did not use a classifier that said yes or no. There was a gray zone. And so they reported the accuracy as 36 out of 38, but the other two were uncertain.

Similarly, with another paper on medulloblastoma, in terms of using ambiguous classifiers, reporting an accuracy of 72%, but it's not really the right way to do it with ambiguous classification. So that's a separate one. So I think I'll stop here, so I have a little something to say next time I see you.

OK, thanks. Any questions right now? So is this-- I'm trying to keep it fairly simple. Does this seem appropriate, not too complex but not too slow either? Or does it seem a little too slow? Whatever, yeah.

AUDIENCE: Is this going to be up on the myCourses website?

STEVEN A. GREENBERG: Yeah, yeah, these are the same ones I used last year. So is the stuff that we put up last year still on there? Does it need to be put up? Yeah, so that's still on there, yep. Have guys been getting problems sets to do? Hoe's that been working?

AUDIENCE: [INAUDIBLE] presentation of basic colors. [INAUDIBLE]

STEVEN A. GREENBERG: There's the Eisen Cluster and TreeView software.

AUDIENCE: OK. Is it TreeView making the graphs?

STEVEN A. GREENBERG: TreeView makes the graph.

AUDIENCE: [INAUDIBLE] if I do something with another tool, then I can [INAUDIBLE]?

STEVEN A. GREENBERG: As long as you put it in the right format for TreeView to look at, yeah.

AUDIENCE: OK, thank you.

STEVEN A. GREENBERG: Yeah.

AUDIENCE: There is another system that used pink and blue, right?

STEVEN A. GREENBERG: There's the pink and blue system too. So some of those are like Peter Park, who's going to talk to you, I think, next week, he just uses our programs that he wrote to do that. There's commercial software called GeneSpring that uses pink and blue. There's a program called TGEV, T-G-E-V, out of the TIGR, The Institute for Genomic Research, and they have free clustering programs.