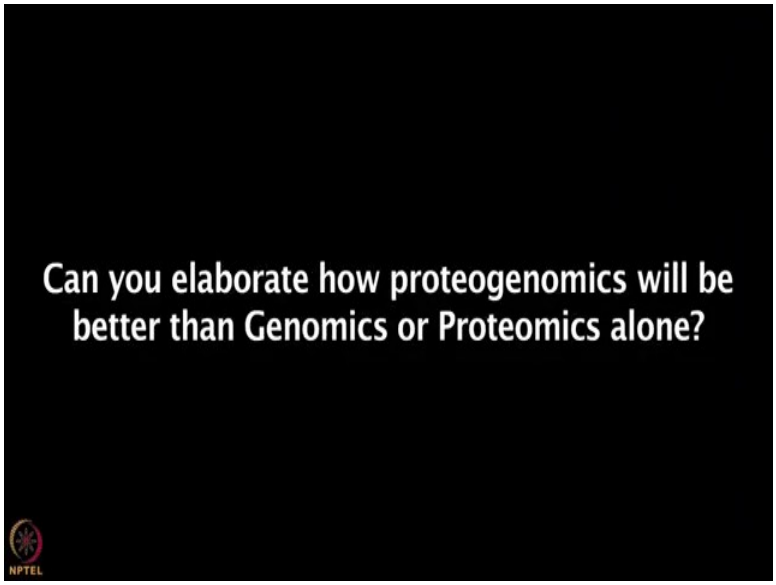
Introduction to Proteogenomics

Dr. Sanjeeva Srivastava
Dr. Henry Rodriguez
Dr. D. R. Mani
Department of Biosciences and Bioengineering
Director, Office of Cancer Clinical Proteomics Research
Principal Computational Scientist
Indian Institute of Technology, Bombay
National Cancer Institute (NCI), National Institutes of Health (NIH), USA
Board Institute of MIT and Harvard, USA

Supplementary - 1
A perspective on Proteogenomics

Hello, my name is Dr. Henry Rodriguez and I am the director of the National Cancer Institute Office of Cancer Clinical Proteomics Research.

(Refer Slide Time: 00:35)



I think the beauty of protogenomics is that it offers the ability to provide a more comprehensive picture of the underlying biology of cancer. And, I think that was already unequivocally demonstrated by the National Cancer Institute CPTAC program; with actually demonstrated the ability when you combine proteomics comprehensively about a comprehensive layer of genomics both in colorectal cancer, breast cancer and ovarian cancer. You are able to pull out additional biology that is either difficult to obtain or simply not feasible through one omics based approach.

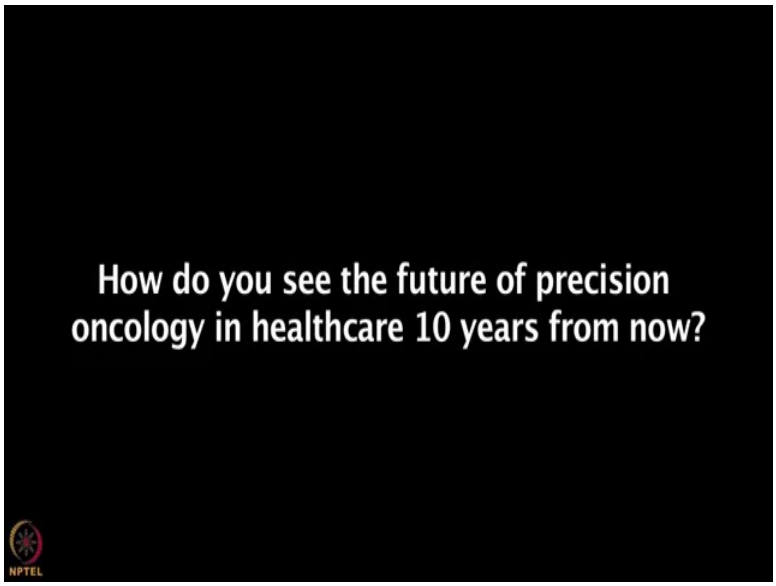
So for me, it is the convergence of these disciplines that begins to shed new light on our ability to not only understand the biology, but hopefully we could translate the biology towards better patient care.

(Refer Slide Time: 01:31)

I would say from a short term perspective, the main one is to take the genomics information and they better connect it to a functional environment. From a long term perspective is that you want to take not just the data but the knowledge you are able to extrapolate from the data and potentially take that biology the additional biology is going to be more comprehensive, and again potentially move it towards better patient care.

The reality is I believe that a protogenomics perspective provides a more systems perspective of the biology itself. And, we would like to do is potentially use that additional fundamental knowledge to better identify what type of treatments to provide our patients. And, also at the same time try them to understand how they would be responding to not only existing therapies, but potentially next generation phase therapies.

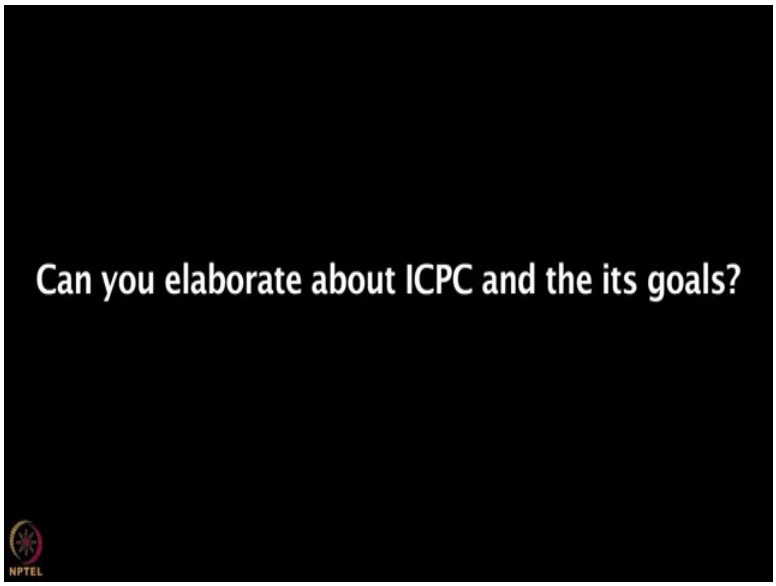
(Refer Slide Time: 02:25)



Actually what can I see more is as technologies are maturing and our ability to measure more things not only at a tissue level, but even at a cellular level; I am firmly a believer that you are going to see more-more a convergence of different disciplines. You will see genomics converging with the proteomics, the proteomic converging along with the genomics any imaging you throw into the mix and even metabolites. So, for me it is the blending of these disciplines that you are going to see more-more over the let us say the horizon of a 10 year window.

But, more importantly is that the information that is being developed the data; the big data as a lot of people refer to it I think that is going to be a key array that is going to be very promising in the years to come. For me quite frankly, I actually see data as the new oil and our ability to be able to not only look at the information, but the ones it is going to be able to actually apply a lot of artificial intelligence deep learning and extrapolate the knowledge; I think that is we are going to see a lot of fundamental breakthroughs when it comes to precision oncology.

(Refer Slide Time: 03:34)



Absolutely; so, ICPC the beauty of that I think that is one of the initiatives that was inspired by the United States Cancer Moonshot effort. Today it is an incredible program; currently it involves 12 countries that spans over 31 institutions collectively. All these institutions in these countries are now working together to try the better understand over a dozen cancer types at the molecular level. I think the part that also makes it extra special is that each one of these institutions in these countries have pledged of the molecular data that they generate, they are going to make it available to the public.

So, for me the idea of ICPC the goal, the ultimate goal is actually quite simple and that is ultimately to develop an international database, that is now is going to be representative of the diversity of people along with their cancers around the world and making that accessible to the rest of the population across the globe.

(Refer Slide Time: 04:43)

You know for me the idea of putting data in the public domain actually sends from three fundamental principles behind it one, I think a lot of the information quite frankly is going to be pre-competitive. But the part that is quite nice about it is that by putting it out in the public domain, it allows other individuals to look at the data sets and hopefully it stimulates new hypotheses along the same cancer that most likely was not hypothesized prior. So, it stimulates new research at a fundamental level.

Secondly, I think by other people getting access to data, if further stimulates the development of new computational tools and we hope that those computational tools are able to identify new discoveries within those original data sets. And quite frankly, what I have noticed is by putting this data in the public domain, you actually bring in new people into the research and you make it a much more multidisciplinary than it prior would existed in the years past. And, by bringing new individuals making it multidisciplinary, bringing computational sciences along with the people that produce the data, I hope you could take the fundamental biology and extrapolate the new knowledge that hopefully will be translated towards cancer care.

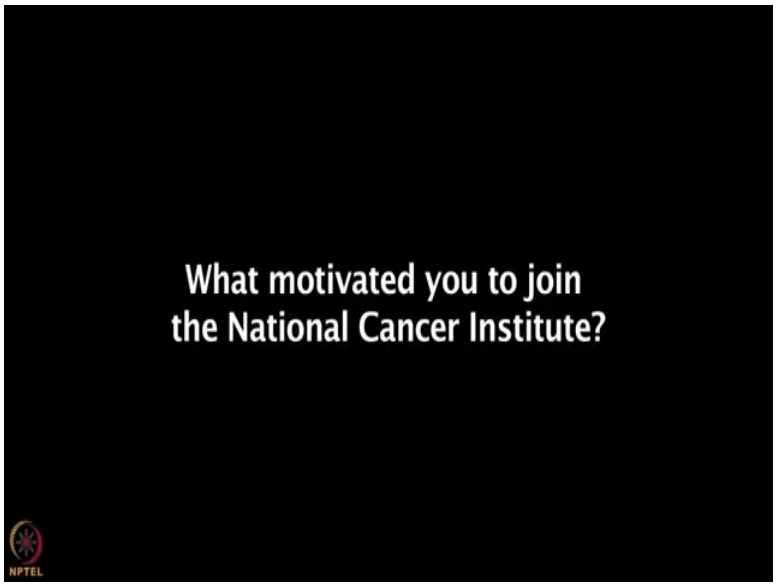
(Refer Slide Time: 06:03)

So, one of the lessons that I have now learned over the past just over 10 years of having the privilege and the honour to kind of lead the National Cancer Institute's Clinical Proteomic Tumour Analysis Consortium is that the fundamental belief. And, truly now the knowledge that by having more disciplinary research groups in the space of oncology, actually accelerate science.

The other component that really inspired me was to call for the US based Cancer Moonshot and its overarching objectives, which are very simplistic: one accelerate cancer research. For me a lot of that involves in developing these team based programs, but the other two which were very key in the cancer moonshot was one and first and foremost is greater cooperation in collaboration amongst researchers not just within one institution within a country, but across countries.

And secondly, is making the information available to the public and that is something I have been very passionate about, we have been doing for over 10 years now. In fact, 15 if you look at genomics landscape of what NCI is done within TCGA now CPTAC and now expanding that to other people across the globe. So, for me the Cancer Moonshot, what it represents is hope; hope that is going to be offered not just for the research community, but also towards patients and their loved members that are afflicted with cancer.

(Refer Slide Time: 07:36)

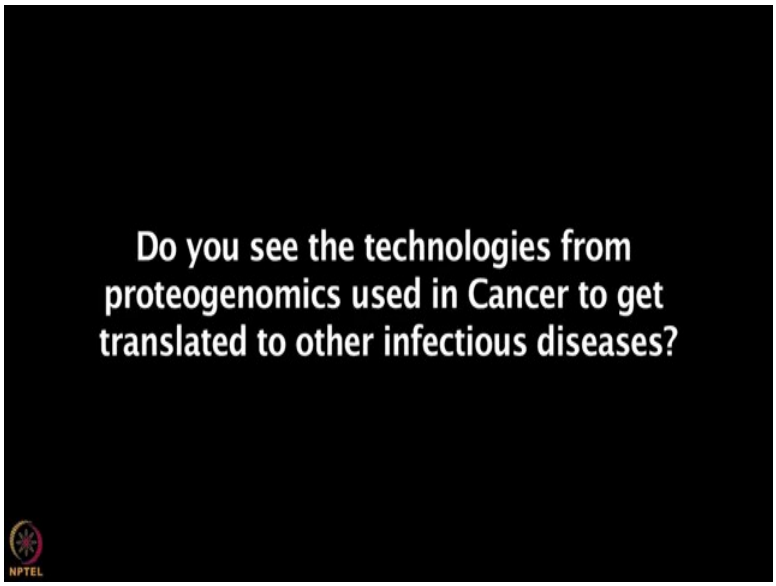


You know one of the things have has always drew me in life is the ability and the willingness to take a risk and I know the NCI when they actually asked me years ago they actually join. One of the things that I have admire about the NCI is that it is an organization that enjoys taking risk and what I am and what I mean by that our two initiatives that are very dear to me and standout. The very first one is the Cancer Genome Atlas that was a big risk for the NCI; we did not know what would come out of it.

But we always had this feeling that by looking at a tumor, looking at cancer at the molecular level that we begin to better understand and ravel the mysteries of nature when it comes to that disease. And, at the same time we took the same risk when it came to proteomics specifically with the CPTAC program.

So, for me one of the driving factors is the willingness to take risk along the same lines now, I think that we have taken with the Cancer Moonshot both in the APOLLO and ICPC. It is that fundamental belief that if we just take that low risk, go and explore an area of science, that we think that there is a glimmer of hope; the belief is that taking the risk you will have rewards and the rewards ultimately for us is to translate it towards better patient care.

(Refer Slide Time: 09:01)

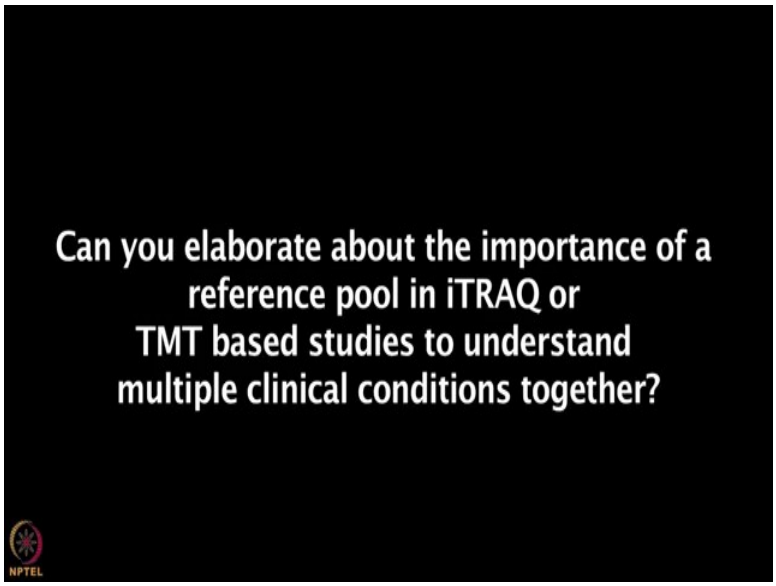


So, the simple answer is absolutely; the reality is technology is technically ambivalent to the biology we are trying to go after. So, if at the fundamental core we are trying to find out is can I identify very key molecular signatures that could better help me understand the disease as a whole than both genomics, proteomics and even the convergence of those two from a proteomic perspective will absolutely be beneficial.

Furthermore you could also look at these different technologies and the disciplines and potentially, then begin to develop diagnostic techniques to be able to detect such infectious diseases and not really within a city, but even in remote villages if that it its going to be even more important. So, I fundamentally do believe and I think a good understanding is technology is really not specific towards a disease that is the beauty, when you develop technology from one discipline it could easily be applicable to another discipline.

So, I am D. R. Mani, I am a principal computational scientist at the Broad Institute of Harvard and MIT. I am in the proteomics platform there and my primary role is to apply statistics and machine learning methods to the analysis of all kinds of proteomics data. So, we look at discovery proteomics, targeted proteomics, proteogenomics we apply computational methods and algorithms to the analysis of all kinds of proteomics data with the hope of achieving a rigorous approach to analyzing data. So, that whatever comes out is defensible.

(Refer Slide Time: 10:57)



So, the main reason for using label proteomics methods is to make sure that you can achieve higher throughput than is currently possible in proteomics. So, right now genomics can do really high throughput you can sequence genomes very fast, but in order to do proteomics and get a proteomic profile for a sample it takes quite a while. And so, it helps to be able to multiplex samples so, that you can hopefully increase throughput by 5 fold or 10 fold in many situations. In order to do that in a large study you really need to be able to run many of these multiplexed experiments and then we able to connect them together.

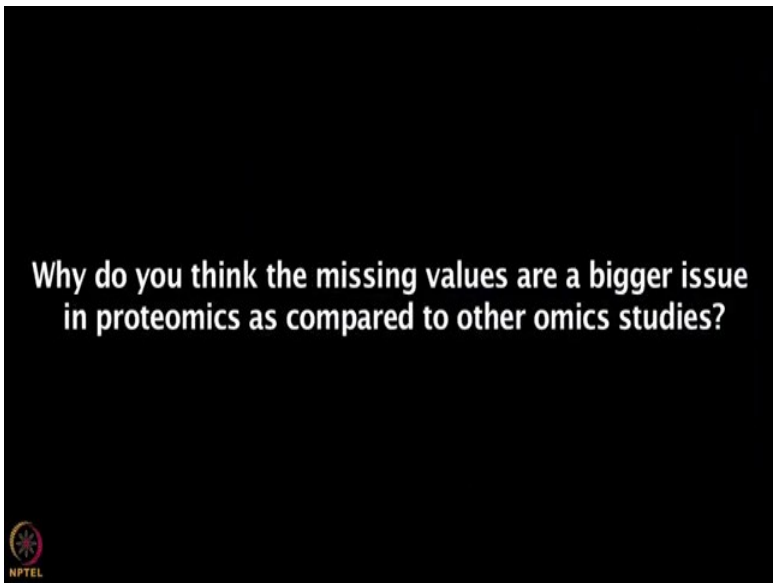
So, if you have 100 of samples in a relatively large study, you would need many different experiments to accommodate all those if you are doing multiplexing 5 or 10 samples in each experiment. And when you do that you have to have some kind of a way to link all the data together so, that you can put all your samples together and then do a statistical analysis.

And so, in order to do that the primary tool that we use is what we call the reference sample, and in most situations this is created by combining a pool of different samples in your project. But it is done in such a way that you either use all your samples or if you are using a subset you sample the subset to represent the groups in your original project.

So, that there is no bias in terms of what went into the reference pool, but once you have that you kind of create a large vat of a sample that you can put on every one of your

multiplex experiments. So, that at the end of the project you can use that pool to read out variation from experiment to experiment in some way normalize on a experiment by experiment basis. So, that you get data that now you can compare across different experiments. So, that all your samples can be put together into one table and then you can perform your statistical or machine learning analysis.

(Refer Slide Time: 13:27)



So, in proteomics missing values are a bigger issue because the proteomics methodology of how you obtain a proteomic profile. So, you inject a sample into the mass spectrometer and then you do what is called data dependent analysis or even if you do not do that. But use other methods there is no reliable way of obtaining a measurement of every protein in your sample with genes. For example, in genomics if you are doing RNA profiling it is more easy to get a catalog of all the transcripts you would like to see, and then put them on a chip or even if you are doing RNA sequencing without microarray or a chip you can still kind of see almost all the genes that are expected to be present.

But with proteomics the issue is that it is a very stochastic, the measurement is a very stochastic process and so, you end up not measuring many of the proteins that are present in your sample. So, in most situations in proteomics if something is not measured it does not mean it was not there, it could also be because it was there. But you were not able to

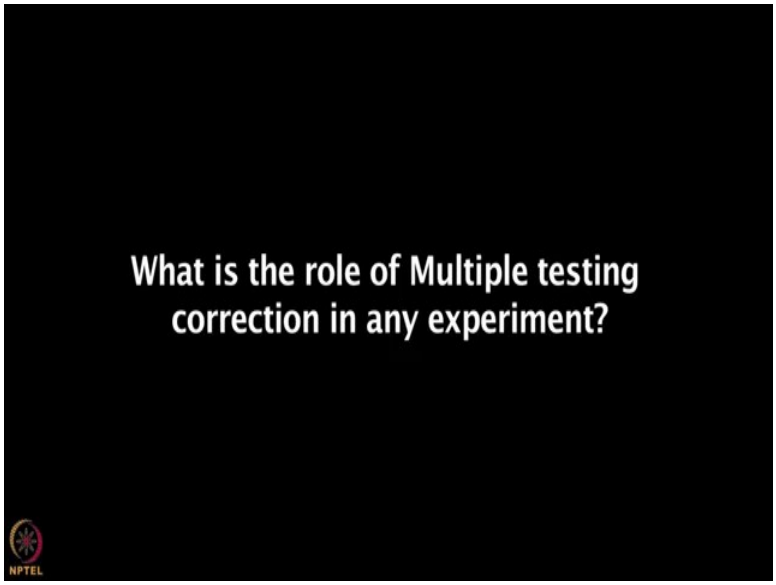
see it with your measurement methodology that is less of a problem with genomics and so, missing values have to be treated more carefully in proteomics.

And if you go to looking, if you start looking at post translational modifications like phosphorylation or acetylation, then the problem is even more compounded because a phosphorous site that might be phosphorylated in one sample may not be phosphorylated in another sample. And, when you are measuring these phospholipids; those phosphor peptides may not be seen in many other samples and so, the missing value problem is much more compounding.

And so, analysis of proteomics data now requires more careful thought on what to do with missing values. There are many ways to approach the problem, but I think the bottom line is that when you are analyzing proteomics data you have to be constantly cognizant of the fact that there are missing values, the fact that these missing values are related to abundance. So, in other words the values are missing because the abundance is most likely low.

And, in those situations statistically you have to be very careful how you deal with missing values. And, in many cases you might want to use tools that can either systematically handle a missing values or if you are going to remove missing values it has to be done in a very careful and thoughtful manner.

(Refer Slide Time: 16:17)



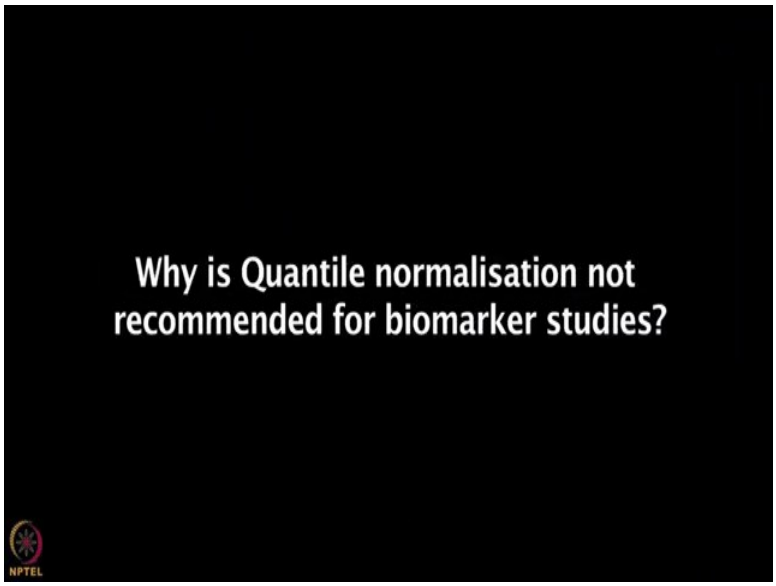
So, when we are talking of proteomics or genomics or proteogenomics we are talking of experiments where a large number of things are measured. So, in genomics you could measure like 15, 18 or 20,000 genes. In proteomics you measure 10 to 15,000 proteins or if you are looking at phosphosites in a study you might have 25, 40 or 50,000 phosphosites you have measured.

And, when you are trying to use this data to find what is differentially expressed in groups of interest for your study like cancer versus normal or different cancer subtypes, you do what are called marker selection or marker analysis. Where you try to find markers that are up regulated or down regulated in subgroups or of the sample set that you are looking at and when you do that you apply standard statistical tests like t-test or f-test or rank test certain and many different tests. And, the problem with these tests is that if you repeatedly apply them on a large number of features in this case genes or proteins.

You can end up with things that appear to be statistically significantly differential in your groups just by random chance. And so, the more tests you do, the more likely it is that something might appear to be statistically different between your groups while it is not really the case in reality. And so, to account for this and to have results that are more robust and kind of more believable from a biological perspective, you would want to apply what is called multiple testing correction.

So, here the statistical significance that is assigned to a test is adjusted because you are doing many many tests like thousands or tens of thousands of tests. So, once you take that into account your statistical significance is reasonably adjusted and then you can get results that are more believable with fewer false positives. Even after that you still have to be careful to make sure that your cognizant that there could be false positives or other false discoveries in your data. But, using multiple testing correction is the first step to kind of getting results that are more robust and are worth following up from a biological perspective.

(Refer Slide Time: 18:52)



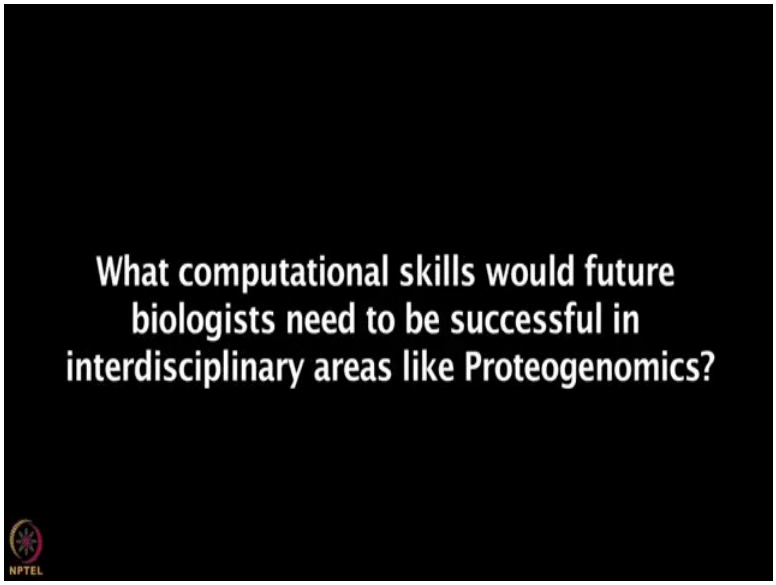
So, most studies whether its proteomic or genomics start off by doing normalization where you are trying to put all the samples on an equal footing. So, that you can compare across samples and analyze them and any minor differences in how the sample was prepared or how much was loaded on the mass spectrometer are kind of subtracted out. So, that what is left is mostly the biological differences between the samples. And one way of doing that kind of normalization is called quantile normalization where, the quantiles of the observed values of the proteins or peptides are kind of made the same; in other words the distributions are marked to kind of make all of them similar.

So, when you do that things that are extreme in other words so, you were doing a cancer normal comparison and there was a protein that was very highly regulated in cancer, but not, but the proteins that are regulated in a normal sample, do not have that extreme values that they have achieved. When you do quantile normalization then it kind of squashes, the signal in the cancer or the extreme subset of samples kind of pulling them together towards the kind of pulling them towards the mean.

And, what that does is your ideal biomarker signal will now be either not strong or could actually get obliterated because of the way you did your normalization. And so, in general quantile normalization is not as approach to use on an average project, in other words it is not the standard approach one would consider using and if it is used, it has to

be used very carefully and you would need to think about why you want to use it and if it makes sense.

(Refer Slide Time: 20:51)



So, I think the way computer science has evolved in the last few years artificial intelligence and machine learning and all those areas that the buzzwords that you hear are becoming actually more and more useful in doing analysis and kind of making sense of large amounts of data in a large number of fields. So, they started in computer science, but now they are almost universally percolating to all other areas.

And, I think in biology as we go into the era of omics with genomics and proteomics and proteogenomic data we are going to collect more and more data. And the biologist who has the domain knowledge of the kind of things they are looking for in their studies will need to be will need to have some knowledge of the kind of tools that can be used and how to correctly apply them.

So, I think the biologists of the future will need to have more significant understanding of what computational tools are available, when they are applicable and to some degree also be able to apply them at least for simple everyday settings where they are generating data. They should be able to analyze their data without having to wait for a computational or a bioinformatics scientist to come and look at their data.

So, on a day-to-day basis I think the future biologists should be comfortable doing their own data analysis and when it comes to special or one off analysis or analysis with more complex study designs then they definitely should collaborate with the computational scientist. But, they should also be in a state where they can understand what the computational scientist is doing, be able to speak their language and to be able to understand whether the techniques applied are appropriate or not.

And, to some degree it also applies in the other direction, computational scientists should also know enough biology to have a common language with so, they can carry on an intelligent conversation with biologists. And, strong collaborations with people who have deep roots in computation and deep roots in biology I think is the future of good research and so, both teams should be able to converse with each other and be able to understand each other's fields a little more than is currently done I think.