Introduction to Proteogenomics

Dr. Sanjeeva Srivastava
Dr. Bing Zhang
Dr. Karsten Krug
Department of Biosciences and Bioengineering
Indian Institute of Technology, Bombay
Baylor College of Medicine
Broad Institute of MIT and Harvard

Supplementary S - 18

A Perspective on Proteogenomics - III

So, my name is Bing Zhang, I am a Professor of Molecular and Human Genetics in the Baylor College of Medicine in Houston Texas in the United States. So, I am a PI of lab of about 10 people. So, our focus is on computational biology based application to studying cancer. So, my lab is supported by funding primarily from the national institute of health and the we were currently applying integrative bioinformatics methods to the understanding of cancer and try to figure out better ways to treat cancer.

(Refer Slide Time: 01:07)

I think the probably two aspects of this; first I think people have to understand the need.

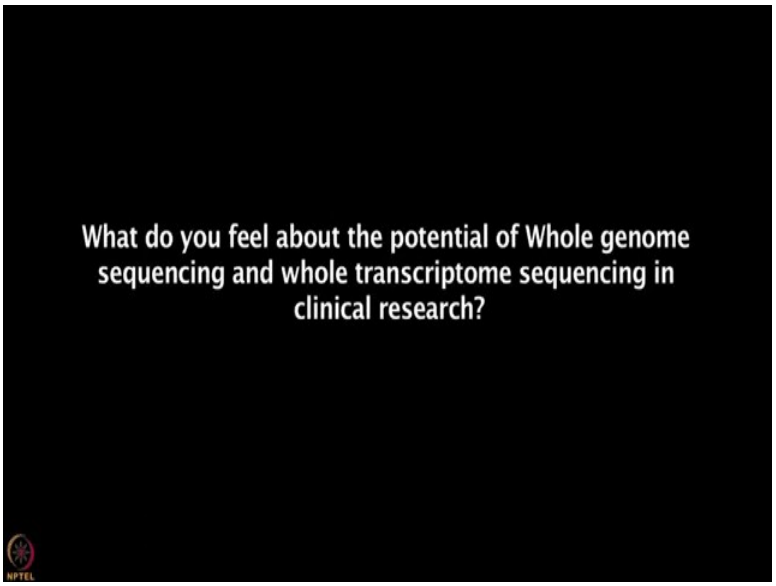
So, why we need to do the integration, I think there should be some demonstration project to demonstrate that there is a value to integrate the data and I think some of the government effort for example; the CPTAC effort in the US is doing something like this to try to bring

people together and the demonstrating the value of by integrating genomic and proteomic data. For example, we can learn something that we would not be able to learn by looking at the data separately, I think then we can get some papers published for example, and then people will appreciate the value of this.

And the second part is, I mean even if you want to integrate you need to have the ability to integrate the data right. I think currently this is through interaction between scientists with different type types of expertise; but I think in the future we also need to think about education. Our next generation scientists should be able to at least understand post genomic data, proteomics data etc and also can understand the computational part and the experimental part.

So, that they have a holistic view of the biology, and then they can do much better job than what we can do today.

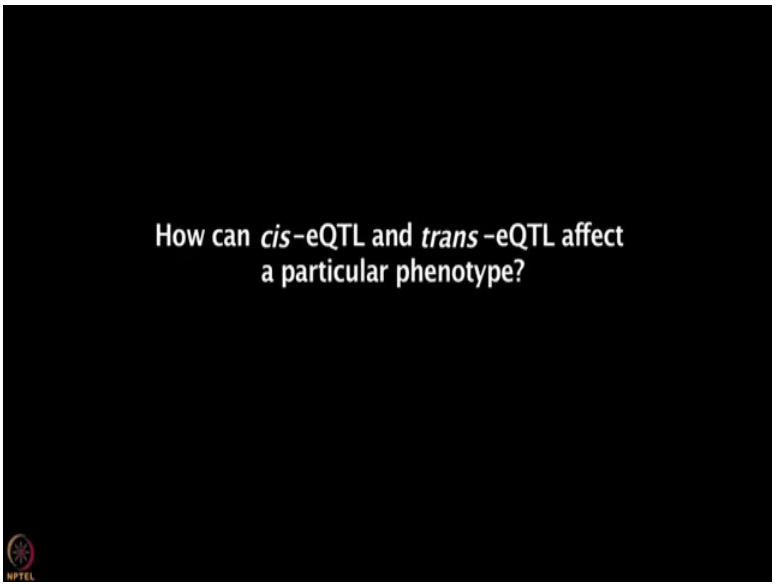
(Refer Slide Time: 02:44)



Well, I think basically these are basic technologies and the whole genome sequencing technology is trying to study the DNA sequences and transcriptomic sequencing are trying to study the transcribed mRNAs. I would rather also throw in the proteomics part, which will study the protein which is a translated product of the genes. People sometimes ask me I mean which one is a better technology, which one do you like to me I think a better way is to integrate all these technologies.

Again I mean we want to get a holistic view of the system and the way to do that is by analyzing the system at different molecular levels if you can and then using informatics approach to integrate this data and this will give us the better view of the system and better understanding of the disease if you want to study certain type of diseases.

(Refer Slide Time: 03:57)



I think the cis-eQTLs and trans-eQTLs they are not exclusive, actually a lot of times they are the same or they are connected.

Let us think about transcription factor, if there is a QTL in the promoter region of the transcription factor, it will be cis- eQTL because the as SNP in that region will affect the expression of the transcription factor itself but because the transcription factor itself alter the expression will change the activity of the transcription factor, and then it will change a lot of downstream genes regulated by the transcription factor, in that way the SNP will also be trans eQTL.

So, and a lot of times and we can use a cis-eQTL to identify the genes that are potentially important but the trans-eQTL can also tell us what is downstream regulatory network that this SNP is working on. So, I think they are actually connected, they could be connected.

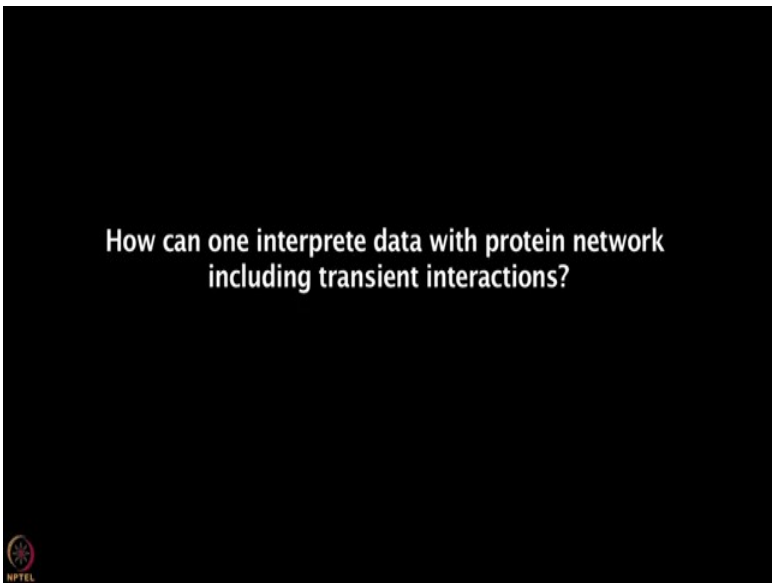
(Refer Slide Time: 05:26)

So, I think that is actually a very fun part of our research. I think bioinformatics as a research area one thing we are doing is to trying to develop algorithms and the methods in order to solve problems.

But eventually what we want to do is to solve the biological problems. We want to make the biologists who do not have the ability to program; for example, to have access to the tools. So, a lot of efforts in my lab have been spent on this direction. So, basically we develop web based applications, so that through very user friendly interface, people can have access to a huge amount of data and then they can also have appropriated tools that they can directly use to analyze those data.

One example is linked omics to be recently published, I think it has been used by started to get many users from the cancer research community as; yeah I think it is really interesting a part of the research to make your tools or methods directly over both of biologists.

(Refer Slide Time: 06:52)

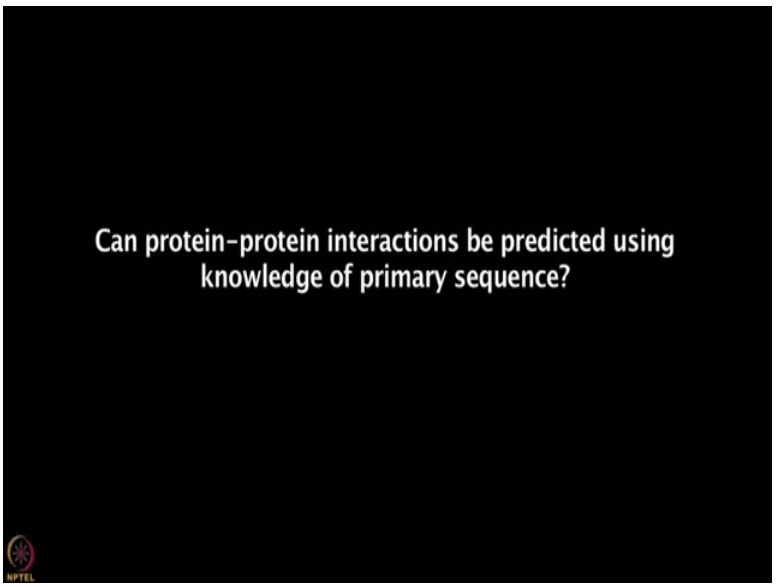


Actually most of the protein interaction network data, we can have all the protein interaction data we can have today in the public data repositories.

I would think most of them are the static or more stable interaction relationship rather than the transient interaction relationships. It is I would hope that more experiments can be done in this area that can help us to identify the transient interactions, and then condition specific interactions; and then we can better annotate the network and the interactions within the network and then we can use the right network to the interpret our data in the right conditions.

So, I think it is not that we already have a lot of transient interactions in the data that; but it is I think we just have very few of those interactions and we need to add more. But of course, we need good annotations in the database to let people know about that, so that you can identify the right interaction network for the specific condition they are interested in.

(Refer Slide Time: 08:16)

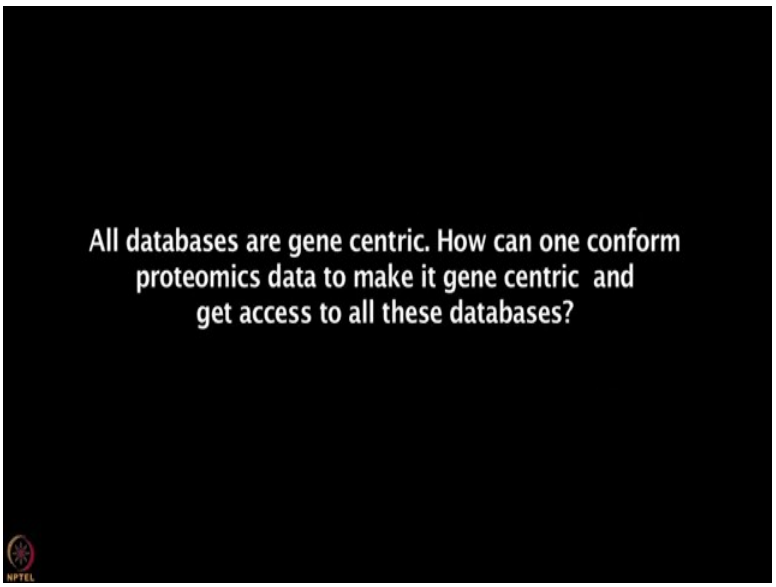


I think the primary protein sequence can provide a lot of information that you can use to predict protein-protein interaction.

But the studies have shown that by only use that information you would not be able to reach where high prediction accuracy. Leveraging other type of data can certainly improve the prediction accuracy and I also want to mention that is technology like the deep learning and this more advanced machine learning technologies that are available today because of the both software and the hardware improvements.

Now, can enable us to better predict the protein interaction for example, based on the primary sequence; but still and if we can incorporate more other type of data that can certainly improve your prediction accuracy and especially when you want to predict the condition specific interaction, I do not think the primary sequence can give you a lot of information on that and for that part specifically you want to incorporate more information.

(Refer Slide Time: 09:38)



That is true on most of the databases even the protein-protein interaction database and the entities in those databases are actually genes.

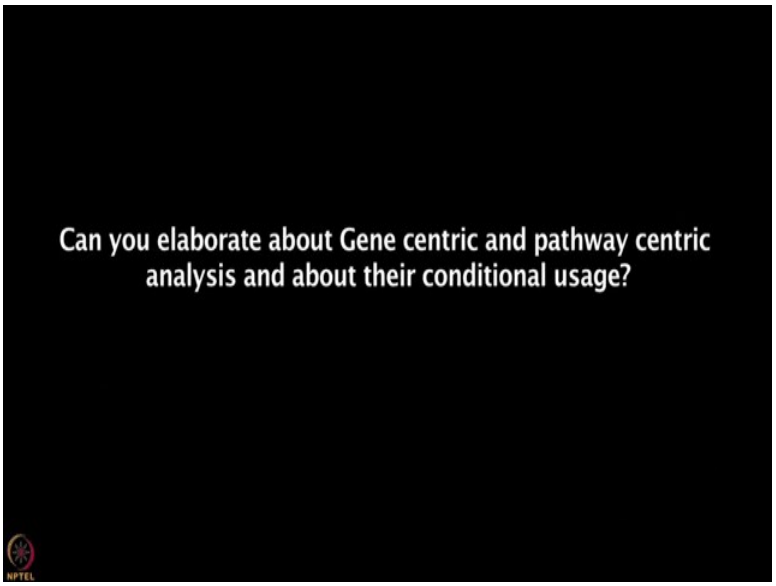
When we talk about protein-protein interaction, the protein interaction database were actually talking about the gene level data and the, I think the challenge is, not the database itself it is gene-centric; I think it we just do not have enough information to distinguish the function and the interaction and the characteristic of the individual protein isoforms. Again I think in the future I hope the protein the databases can be protein isoforms centric.

Because different isoform can actually have very different functions ah; but in order to achieve that and for example, the proteomics experiments, the sequence coverage has to be improved a lot. Because currently if you do a mass spec based the experiment, the sequence coverage is actually pretty low; it is less than 10 percent could be, and with that you would not be able to very well distinguish different protein isoforms that is why the interpretation is usually done as at the gene level.

It is not difficult to convert the protein level data to gene level because it is aggregated to the genome all right. But I think it is more difficult to get the detailed data at the protein isoforms level and impute database centered around protein isoforms rather than genes.

My name is Karsten Krug I am a Computational Scientist at the Proteomics Platform of the Board Institute of MIT and Harvard and I am interested in how we can integrate large scale data sets that have been acquired using different omics technologies.

(Refer Slide Time: 11:56)



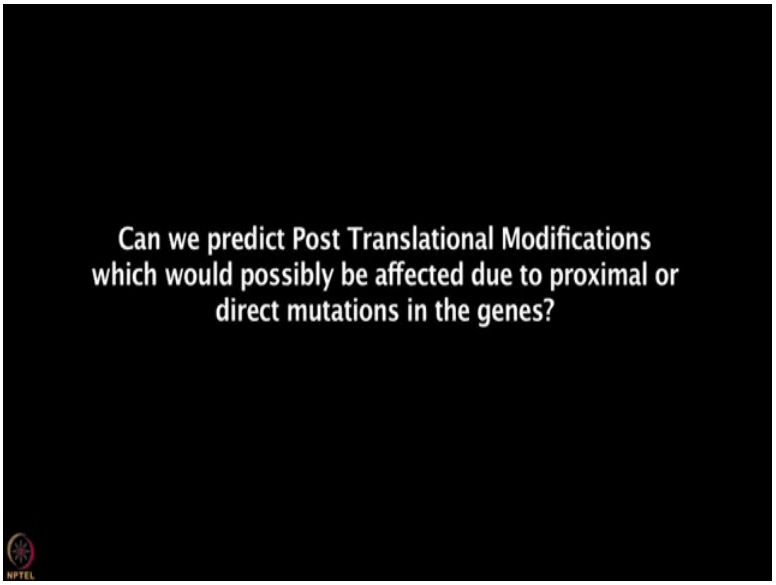
In a gene centric analysis we studied the gene or the gene product itself. So, meaning we want to compare it is an expression between two phenotypes. Let us say in a context of cancer, we want to know whether gene is specifically a up regulated in a cancer compared to a normal tissue which would probably or potentially introduce a new target for this for the specific cancer. If we study the entirety of all genes or gene products in the cell we have to perform a statistical test, which tells us which genes or proteins are statistically significant between tumor normal samples.

And we would end up with long lists of differentially expressed genes, which are sometimes very difficult to interpret. So, in order to better understand what is happening and for example, tumors on the molecular level we would usually or typically map these proteins or genes to pathways in order to better understand; what is dysregulated in these tumors on a molecular level. Yes, many different or several different databases that facilitate this kind of analysis so there is the REACTOME database or the KEGG database, or so the database of molecular signatures or MSigDB.

So, if you ask me whether gene central analysis or pathway central analysis is better. So, I personally think that both types of analysis are equally important. So, sometimes gene centric

analysis and own will probably cannot give you the correct answer, because the gene that you are interested in is probably is not necessarily statistically significant or it is only like a marginal case. But if you look at specific pathways and several members of these pathways are going into the same direction in the tumor sample; for example which, so this gives you more evidence that this pathway is regulated for example,

(Refer Slide Time: 14:08)



So, we know that many many mutations, millions of mutations have been associated to certain diseases and phenotypes; but only for a very few we know actually the molecular consequences that are being introduced by these mutations. So, if a mutation effects coding region of gene, so it might be a non synonymous meaning it can introduce an amino as a change in the corresponding protein sequence and if we think about post translational modifications like phosphorylation of serine threonine and tyrosines this can actually affect these phosphorylation sites.

So, serines and tyrosines these can actually very abundant in the human proteome and therefore, it is very likely that these amino acids are affected by mutations. So, they were probably the simplest case of like the impact of mutation of on phosphorylation sites is that, a phosphorylation site now gets mutated into a different amino acid.

So, meaning it cannot be phosphorylated anymore and we, so it is very crucial to understand what kind of downstream effects these kind of kinds of events have. Like in case a mutation effects a modified amino acid like a serine that is usually phosphorylated. Now the serine is

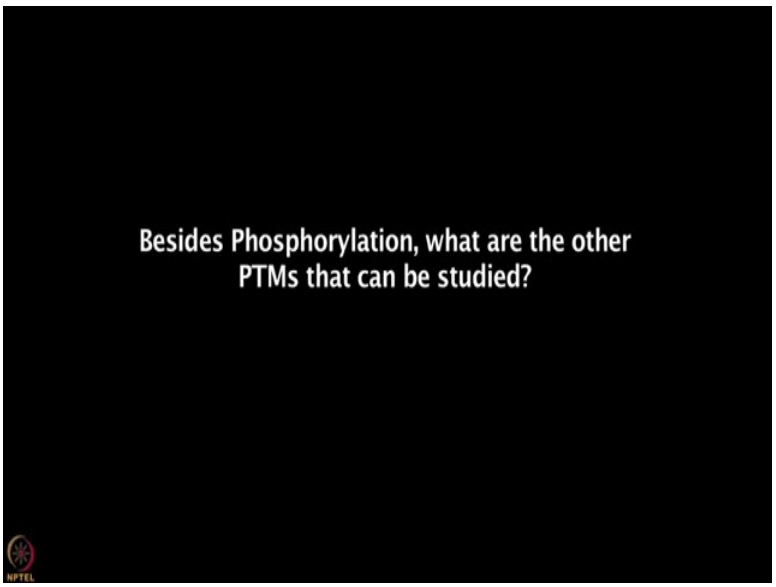
being mutated into a different amino acid, it cannot be phosphorylated anymore. So, and it would it is very crucial to understand what kind of downstream effects are introduced by these kinds of mutations. So, this is probably the most simplest form of such events, other forms of these events can must not necessarily directly the affect the PTM site, but they can be happening in very close proximity.

And for example, in phosphorylation we know, we very well know the enzymes that are responsible for phosphorylation kinases and phosphatases. Kinases which are phosphorylating their substrates are recognizing a very specific stretch of amino acid that surrounds the PTM site. So, this is one of many mechanisms, how a kinase recognizes it is substrates. So, the kinase usually has between a couple and hundred substrate and so these amino acid stretch around these PTM sites is one mechanism how a kinase recognizes it is substrate.

So, if a mutation now changes, the amino acid composition of these flanking sequences as we call them around his PTM site has direct effect on the kinase substrate binding specificity. So, it might happen that a phosphosite has been more phosphorylated by a specific kinase like AKT1 for example, and can now not be phosphorylated anymore by this particular kinase, because it cannot recognize it is substrate site anymore.

So, in the other hand or like another more complex example would be, if the kinase recognition motif now changes from a kinase A to kinase B. So, the wild-type form, the unmutated form, the phosphosite was phosphorylated by a certain kinase and now after the mutation the kinase recognition motif fits better to another kinase, which now can go and phosphorylate this phosphorylation site. So, all of these events are probably not well understood as of now and I think it is very important to learn and to study these kind of events more in detail.

(Refer Slide Time: 17:54)



I think phosphorylation is by far is the best and most studied post translation modifications to date, because we have two methods to study these phosphorylation on a large scale. There is other phosphorylation modifications like your phosphorylation or lysine acetylation which now we also have two methods to study those at large scale or some patient samples.

Of course, we can very easily study whether mutation effects directly these PTM sites, like whether these lysines are being replaced by another amino acid and now these lysines cannot be ubiquitinated or acetylated anymore. But I think we still have very limited knowledge about specific binding motifs for a lot of these acetyl transferases for example.

So, there are specific examples that where we know the sequence motif when we talk about histone modifications for example but our knowledge is still very limited in this regard.

(Refer Slide Time: 19:06)

I think what is very important for a biologist is to be able to at least partially analyze their own data.

So, now biology has moved away from you know hypothesis driven a very targeted type of analysis or experiments more to like a data driven, omics type of experiment. So, the demand of data is on a completely different scale compared to 10 years or 15 years back. So, even as a wet lab biologist it is very important to be able to analyze your own data. So, you need to have some computational skills. I think an easy way I think ok, I think an easy way to get started with any computational analysis or data science driven analysis or scripting languages like R or Python for example.

So, both of these languages are very popular and very heavily used in data science and in general, but also in computational biology in particular. I can only highly recommend any student who studies biology to get some skill set in R or Python.