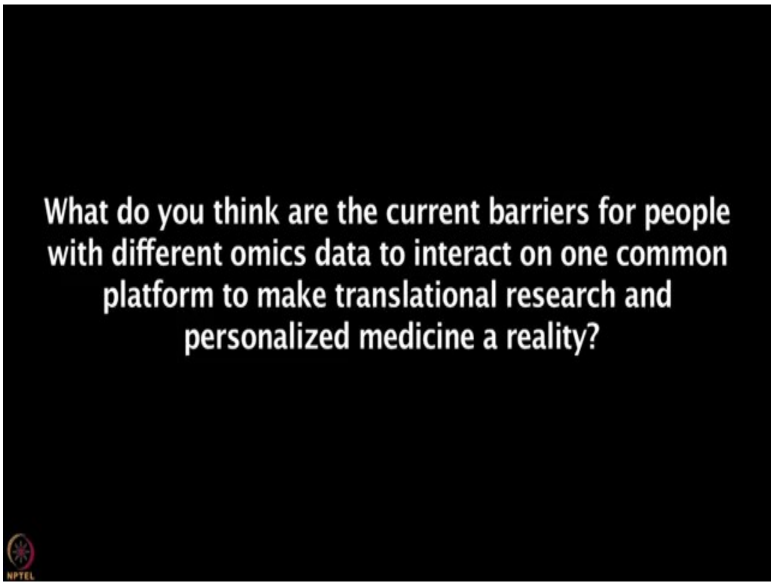
Introduction to Proteogenomics

Dr. Sanjeeva Srivastava
Dr. David Campbell
Dr. Luis Mendoza
Department of Bioscience and Bioengineering
Indian Institute of Technology, Bombay
Institute for Systems Biology

Supplementary Lecture – S17
A Perspective on Proteogenomics – II

My name is David Campbell I work for the Institute for Systems Biology in the lab of Dr. Rob Moritz, we do a variety of proteomics techniques there I work in the lab with Dr. Rob Moritz. The lab does a number of proteomics techniques I myself I am a software engineer. And, so, I work on a number of projects such as peptide atlas and the TPP.

(Refer Slide Time: 00:50)

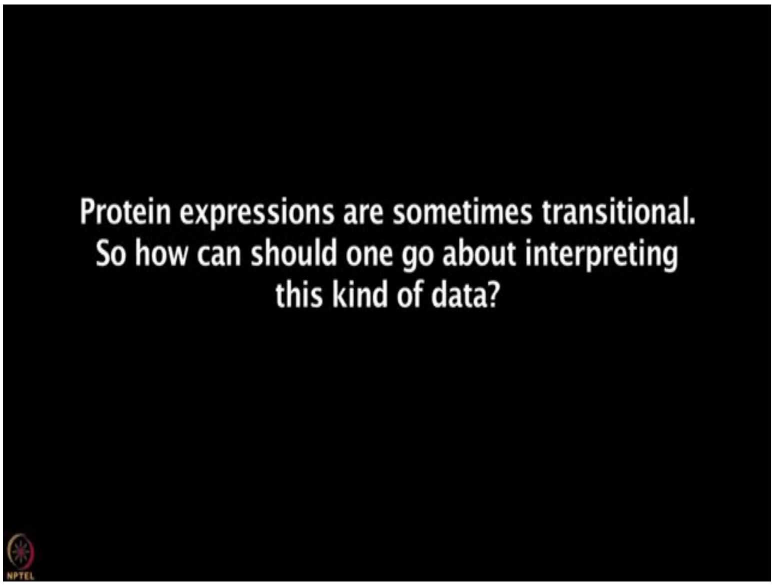


What do you think are the current barriers for people with different omics data to interact on one common platform to make translational research and personalized medicine a reality?

I think there are several different barriers that need to be overcome first of which there are many different file formats all the different data types produce their own distinct file types, they have different fields. So, even within the proteomics field there are many different data types for protein expression. And, therefore, merging those disparate data types can be difficult you essentially need a translator between the different types.

Other barriers are that the various instrumentation can be expensive, the knowledge required to successfully analyze the data, takes some time to obtain. And, so, it is difficult for any lab to have all the expertise and the instrumentation needed to do these types of analysis.

(Refer Slide Time: 01:53)

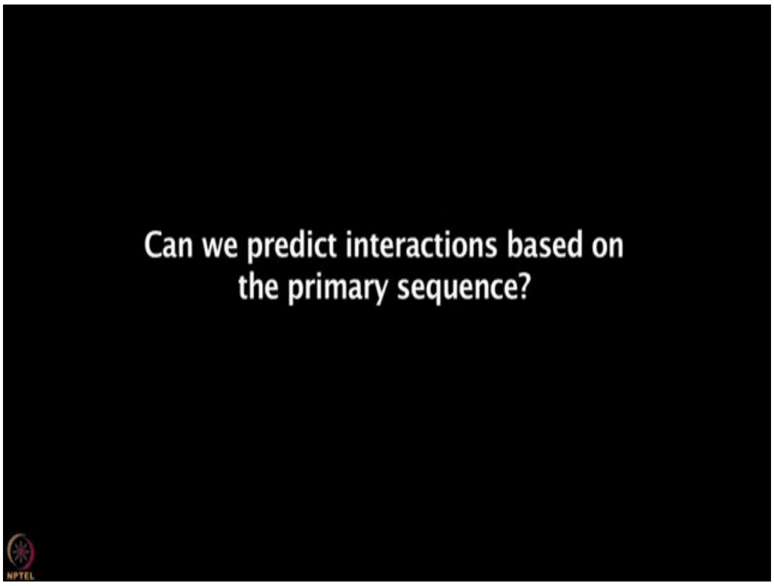


Protein expressions are sometimes transitional. So how can should one go about interpreting this kind of data?

So, protein levels do fluctuate with cell cycle and disease state and things like that, the sort of more transient fluctuations say with cell cycle, should be averaged out if you sample say a population of cells over a whole tissue. This may change in the future when we actually do proteomics on single cells.

But, if you are interested in such cycle related transient differences, you can basically start with a synchronized cell population and do a time course experiment; for instance to see that the rise and fall. But, basically it is just good to keep in mind that these differences do occur and include this in your interpretation.

(Refer Slide Time: 02:46)

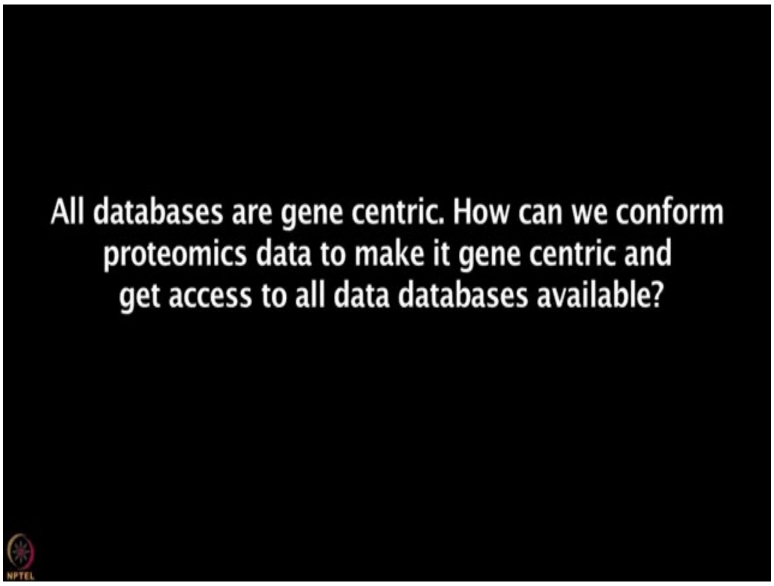


Can we predict interactions based on the primary sequence?

I think this is very difficult to do, because primary sequence with a primary sequence, it is difficult to tell exactly how it will fold and once folded what the characteristics of that particular charge shape, charge space is.

You can approximate this by looking for conserved regions and homology to existing proteins or sub sequences that have been solved. So, you can approximate it, but never really *de novo*.

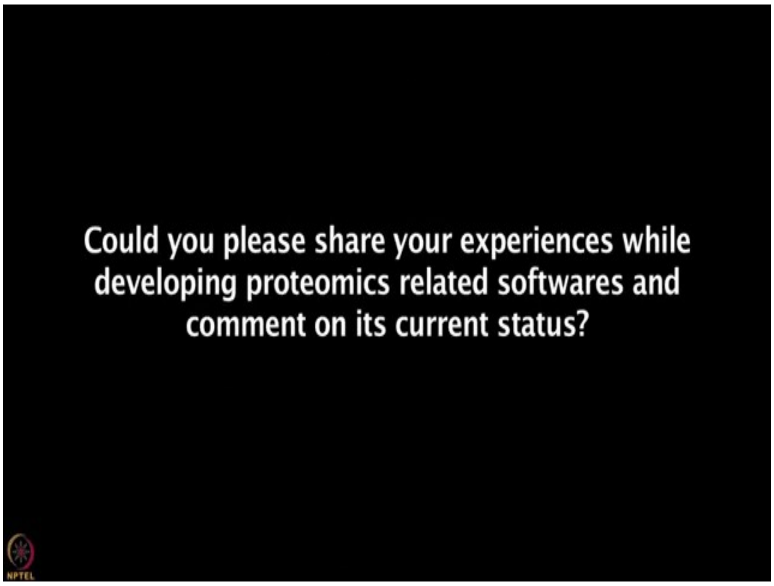
(Refer Slide Time: 03:25)



All databases are gene centric. How can we conform proteomics data to make it gene centric and get access to all data databases available?

I guess ultimately I think that proteomics data can be thought of as gene centric as well, because every protein is the product of some gene. And, possibly some post translational modifications or splicing. So, yes proteins all come from genes. And so, therefore, there is a one to one mapping, one difficulty is that there are different accession spaces between genetics and proteomics. And so, coming up with a robust and reliable way to translate these back and forth is useful.

(Refer Slide Time: 04:08)



Could you please share your experiences while developing proteomics related softwares and comment on its current status?

So, the main projects I have been involved with at ISB, are the peptide atlas and the TPP more so, the peptide atlas, they are both evolving in response to user demand. But, I think that is something very important in general in a software development is that a lot of times software developers want to develop what they think will be a good solution, without really contemplating without really getting feedback all the time from the users. So, I think that is the most important thing is to listen and get feedback and develop what is necessary, not what you think you want to develop.

So, the TPP is pretty mature in the context of data dependent analysis, but we are expanding it in the realm of DIA and other techniques we are continuing educational outreach having TPP courses literally all over the world. In the last 2 years we have had courses in Brazil, in Ireland, in Taiwan, in India and other places several in the United States. So, it really is an educational effort to basically help people understand, what the tool is and how to use it.

(Refer Slide Time: 05:35)

What were the challenges you faced while developing Peptide Atlas project and associated sites?

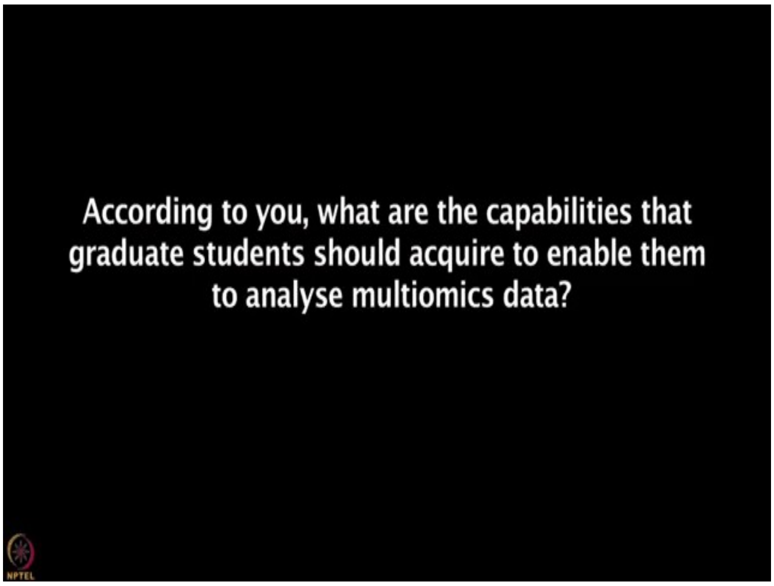
So, one of the main challenges with peptide atlas is that it depends on public datasets. And, so, as soon as we make a build, it is almost obsolete, it is time to go collect more data, reprocess it and remake the atlas. And, we are talking about pretty vast amounts of data, there is a scientist at ISB named Zee San that does most of this data wrangling. And, she is very good at basically processing large amounts of data in a consistent and efficient way. One interesting technical issue from a programming standpoint is the protein super group issue.

So, basically one of the problems I meant to lead to earlier between proteomics and genomics is with proteomics you typically get peptide sequence and peptides can map to multiple proteins. So, it is sometimes difficult to infer from just the peptide sequence exactly which proteins you have. So, there is a program called protein PROPHET and there is other ones, that basically solved this protein inference problem, you have identified these peptides.

So, what proteins, the minimal group that is explained by all your experimental data; so, it turns out protein PROPHET is good at doing this by applying Occam's razor, which is basically the simplest explanation is the correct one. In the peptide atlas there are so, many peptides that we have gotten what is called a super group. So, basically there is enough there are peptides that map to multiple proteins.

And, once and they sort of tie together these different groups and so, it is a little bit hard to explain, but basically because of the sequence homology, we end up and the massive coverage in the peptide atlas. We end up with this huge group of proteins, which we think we have seen, but it is difficult for the existing tools to de convolute and decide exactly what we have seen.

(Refer Slide Time: 07:48)

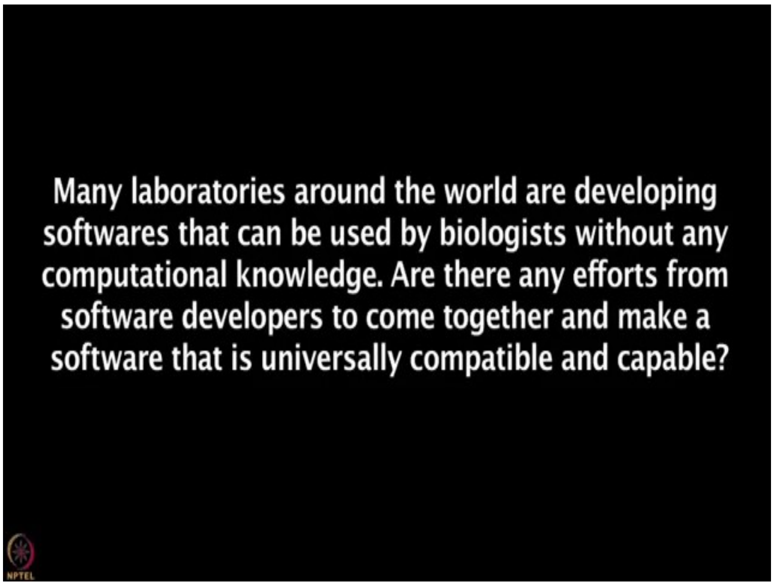


According to you, what are the capabilities that graduate students should acquire to enable them to analyse multiomics data?

I think the most important thing is to take the time to learn about the data technique and the tools available, before jumping into your analysis. It is too easy to you know be eager to push your project forward and basically you know order some sequencing to be done or what have you and then start running tools without really understanding them. The other thing to do is to read the literature and see how other people that have had this have analyzed the same data are approaching this problem.

And, finally, one of the most powerful ways for doing any data analysis is from the command line. It allows you to string together pipelines of programs as we heard discussed today; it allows you to do your analysis on the cloud which is very scalable. So, learning to use programs at sort of the expert level and this is generally the command line, I think is very useful especially for a graduate student.

(Refer Slide Time: 08:59)



Many laboratories around the world are developing softwares that can be used by biologists without any computational knowledge. Are there any efforts from software developers to come together and make a software that is universally compatible and capable?

When you first started the question I thought you were talking about commercial packages, which basically purport, which claim that they can knit together these different data types. And, I think often they can of for very specific data types. I think in the open source community which TPP is part of and there is others there is genomics open projects like SAM tools and others there are efforts to make a common data language.

So, if you have a common data language, then pretty much any tool that you use can be shared or can be extended to use other types of data. So, I think in the open source space, there is a desire and a recognition that interoperability is important. I think from a commercial company they are more interested in having you buy their software. And, so, they are less interested in making everything inter-convertible.

Another scientist at ISB named Eric Deutsch does a lot of work on various standards initiatives, like the proteome standards initiative. And, so, basically they come up with defined file types and defined ontology descriptive languages for communicating information. And, and I think yeah that sort of coming up with common formats and language is the most important part of interoperability.

My name is Luis Mendoza I am a Software Engineer at the Institute for Systems Biology in the proteomics lab of Dr. Robert Moritz. I have been working there for a past almost 15 years. My main goal has been to develop the software for the trans proteomic pipeline, which is an open source free collection of analysis tools, the validation quantification.

And, integration tools a visualization that enable the advanced analysis of a high throughput proteomic data from many kinds of instruments under many conditions. And, we have had great success, we have 1000s of people, over 100s of labs around the world from small labs all the way to big pharmaceutical companies that to some extent use our software and we keep and so, that is what I have been doing there for the last 15 years.

(Refer Slide Time: 11:45)

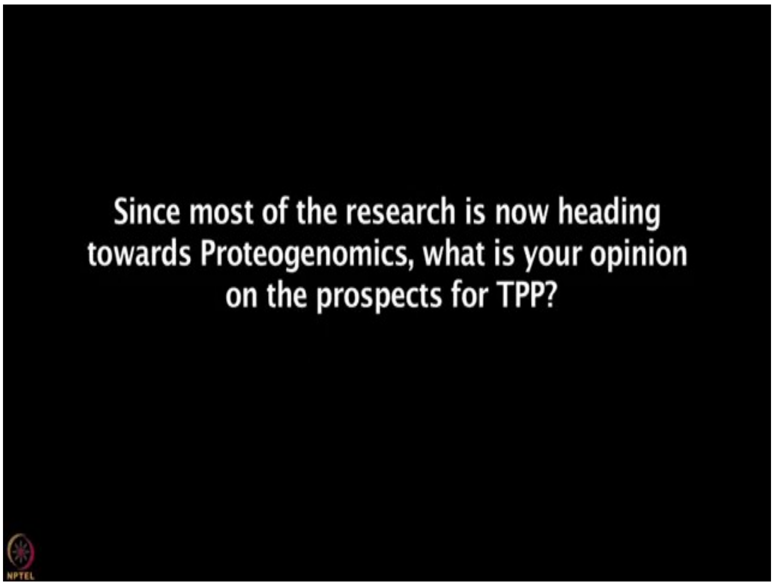
What according you are the current barriers for people with different omics data to interact on one common platform to make translational research and personalised medicine a reality?

Well, there are many barriers I see it one of them is part of the barriers are just purely technical. Obviously, there is all kinds of data being acquired in different omics platforms, but bringing all these data together is still kind of a challenge mostly, because it starts with the researchers many researchers specialize in one area or another and so, when they try to integrate their data there is no easy way to do it.

So, it is up maybe to software to enable maybe doing these connecting genes to proteins to transcriptome data and many other data. So, I think at the moment just having a being able to provide a good software platform or even portals, where what can integrate the data and this has been is already being done to some extent. But that is probably the biggest barrier. Also, learning about the different data and how to interpret different data, genomicist may not really be able to understand proteomic data for example, very well.

So, that is also a common barrier if you do not have a good collaboration with someone else that may make it a little bit more difficult.

(Refer Slide Time: 13:15)

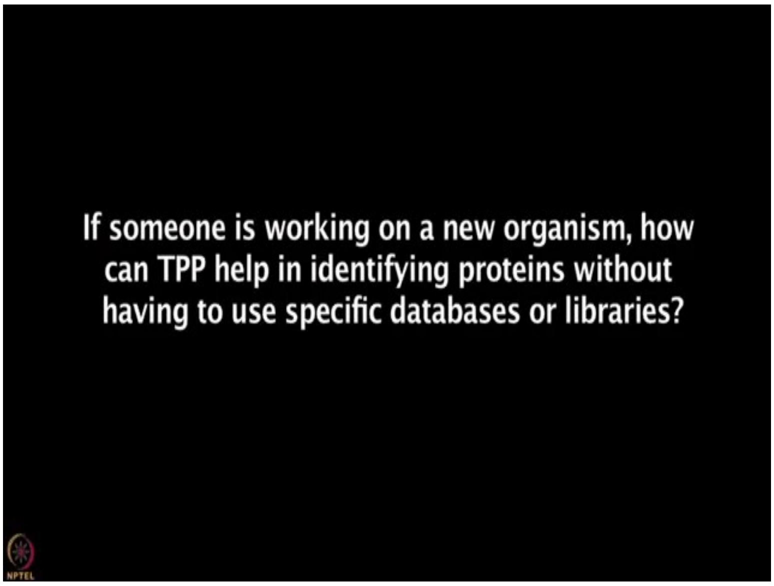


Since most of the research is now heading towards Proteogenomics, what is your opinion on the prospects for TPP?

For TPP specifically for our software even though yes definitely a lot of efforts that are being done to move towards proteogenomics. There will still always be a need to some extent to do the identification and validation of just pure peptides and proteins. So, that that part of it might not go away; with these new techniques over the years we have so, far proven that TPP has been able to evolve to accommodate different kinds of datasets, different kinds of techniques and analysis we are doing RNAseq already we are doing other things.

So, there is very much a possibility that if this is where the field is moving that we will expand our tool sets that expand create new software. If this is why is yeah required for the field that we will be able to provide those or in some case perhaps have third party tools that we integrate as we already do, to have a more complete solution in a single perhaps software platform that people can rely on.

(Refer Slide Time: 14:28)



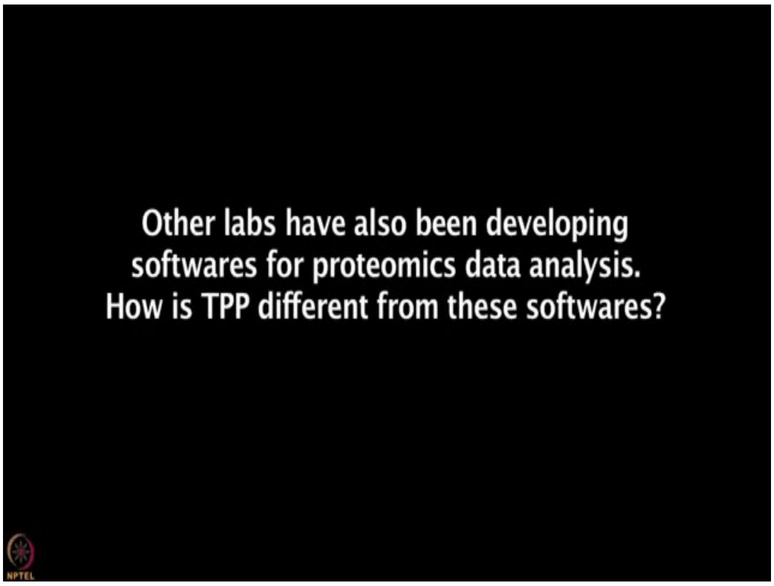
If someone is working on a new organism, how can TPP help in identifying proteins without having to use specific databases or libraries?

This is always a tricky question and, we actually have many users who use TPP for organisms that have poorly characterized proteomes. It is; obviously, indeed fairly difficult to do TPP specifically does require some sort of a reference, this is the basis of just sequence they are researching, but there are other tools out there, that we do not specifically have within TPP other than doing a simple gene translation to just to protein, that allow you to generate a customized database.

For example, you can have you have a RNA seq data, you can from your organism that you are studying, you can generate from that using a set of tools pipeline that is not something we have developed. But, that we are using at ISB to then generate a customized database, that very much looks like a sequence database that you will use to then search against. So, at the moment and the reason we do not have these in TPP yet is for two reasons: one of them other groups have already written these tools.

So, why write the tool again, but the second most important one is because at the moment this data requires a very large amount of processing power and memory and time, and most normal computers are not able to even do this in several days time. We have access to large computers and even then it takes easily one full 24 hour day to even analyze. So, we are trying to figure out ways that we can make this a little more efficient and faster and we are still working on that.

(Refer Slide Time: 16:24)



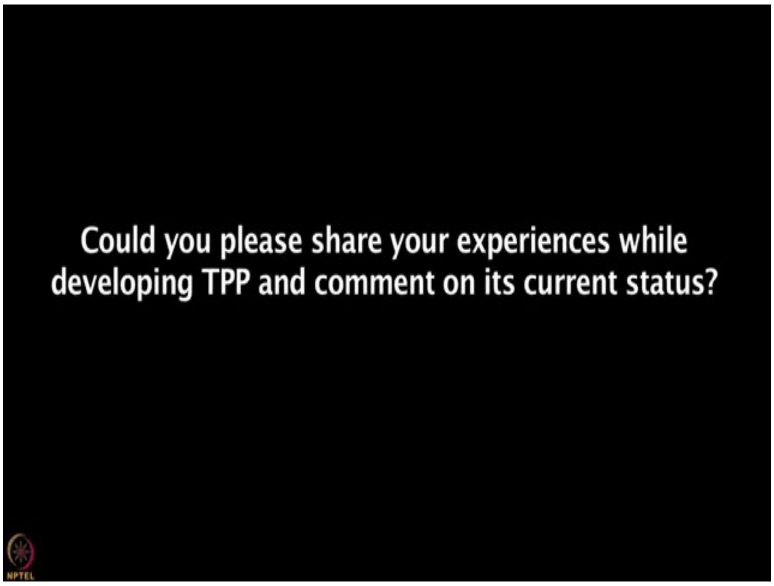
Other labs have also been developing softwares for proteomics data analysis. How is TPP different from these softwares?

I think in general in science you want a combination of collaboration and competition. So, there are other teams that develop certain software that in some ways maybe the community or even we think is better than something we have or maybe something we do not have, and we are able to integrate it with ours to make the whole platform better. Obviously, there are other teams that make similar software to TPP that allows us to have a little bit of competition or lot of competition. I think that keeps us all providing a better software product or even if it is free to everyone.

And so, I do not think it is necessarily one is better than the other a tool, it is as good as your you know your ability to use it. So, you know even a very fast car, if you do not know how to drive it is of no good to you. So, I think for a large amount all of the tools that are still out there that are popular of score perhaps fairly evenly, in the end is really up to the researchers to figure out in their hands, they can use it to get to their to their answers.

And, if it is easily available on you know to a large to a large audience; obviously, we think our tools well are worth a look. Since, we they work very well for us and we often compare them to the other ones and we think they are very competitive and in our hands. In our hands they do perform the best at least in the free software environment.

(Refer Slide Time: 18:13)



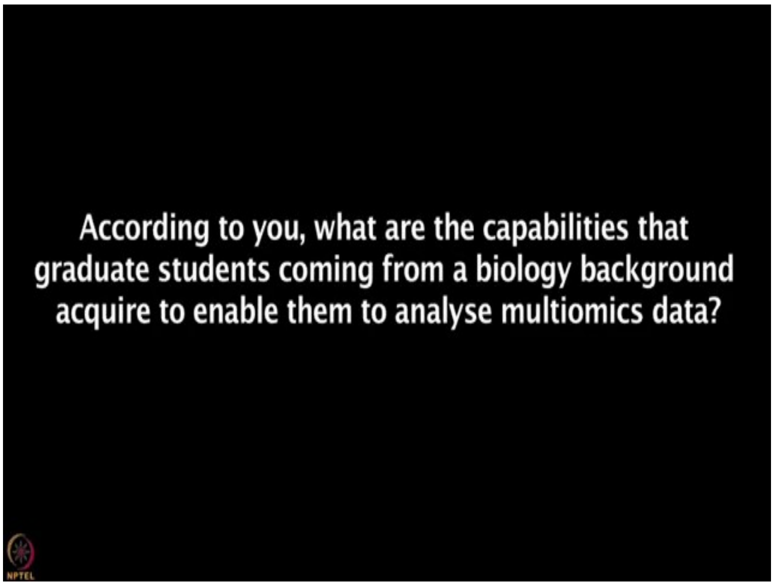
Could you please share your experiences while developing TPP and comment on its current status?

So, TPP is unlike some other software out there it is great because it has been evolving over the last more than 15 years. What started as just a simple program to perhaps validate just peptide assignments has grown into a number of software, that can do end to end all kinds of analysis validation, quantification, different methods, visualization in different ways alignments. And, so, it has been great to be able to provide the researchers mostly in our lab initially and then of course, to the entire or a large part of the community in worldwide community, tools that will enable them to do great things.

I guess as a software developer I feel like very small part of someone else's success when they actually do something very interesting with our tools. And, that is where the true value lies in the tools, it is not exactly you know something I do, but you know that enables researchers to find cure for disease. Or, new ways to you know maybe eradicate some other, you know organisms that you know that are affecting, you know some type of virus or something.

So, in order to make it relevant we over the years we have had to collaborate with our scientists and external scientists and being out here to even when you for example, reach out and teach or present the conferences we get a lot of feedback and that enables the tools to become more mature and perhaps more useful to everyone.

(Refer Slide Time: 20:04)



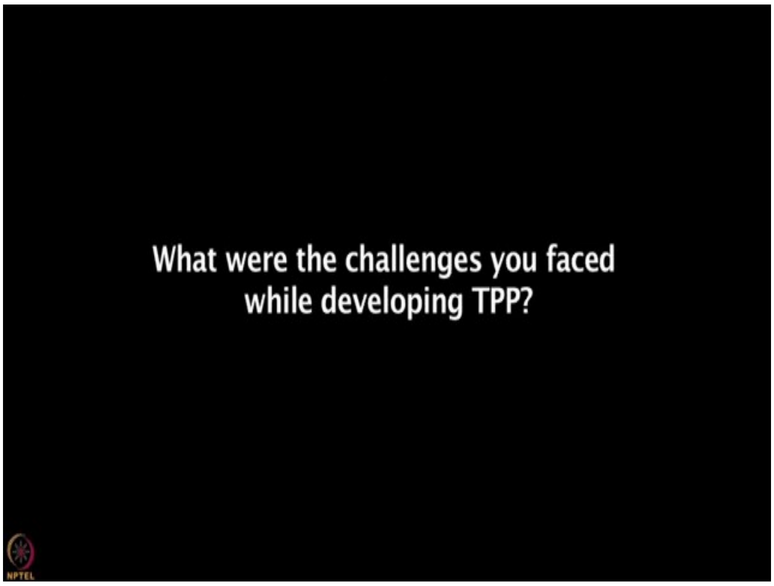
According to you, what are the capabilities that graduate students coming from a biology background acquire to enable them to analyse multiomics data?

Some of the capabilities students need to do this is definitely a familiarity which is with because we have now a high throughput very large data sets, you have to be able to have a basic understanding of just basic statistics in the experiment design. But, also you know basics of how to use several computer software and be able to analyze there or evaluate the results to see which one works best for you. At the most basic level most of the or many of the software's out there will have a fairly easy to use interface.

However, there can be many ways; many reasons why it can be a little bit difficult to use if they have different, there are formats so, you have to familiarize yourself with those pipelines. It is also very useful for students, if they can learn even a little bit to just use things on the command line and other things like R that the statistical language R, because then, you can really unlock other features, that may not be obvious just on a simple graphical user interface. Especially if you are doing high throughput studies with many many samples this makes your life far more efficient and easier than doing this.

So, there are little things that you can do that were perhaps won't be too difficult to learn that will really give you a lot of a value for your time. In northern that; obviously, talk to other students, talk to your professors, talk to maybe people that develop tools to help you figure out how to best use them and how to get to the results that you are looking for. So, that you spend more time doing interesting research than just trying to run some software.

(Refer Slide Time: 22:00)



What were the challenges you faced while developing TPP?

I mean the challenges are still around. So, there is always a challenge and you know like they say you always consider that an opportunity. So, you know there is always times when something does not work, gives you or as new data set that gives you now a strange results that you were not expecting because, you do not have that data set before. And, so, these always come up constantly and that makes the tools more robust.

So, we were able to solve that problem then; that means that after that anyone with this type of data will hopefully have that problem solved for them with the tools. So, in you know so, one of the challenges is just not having all the data around and yeah it is, but that makes it fun for us too and I try to solve the problem and provide an answer.