An Introduction to Proteogenomics

Dr. Sanjeeva Srivastava
Department of Biosciences and Bioengineering
Indian Institute of Technology, Bombay

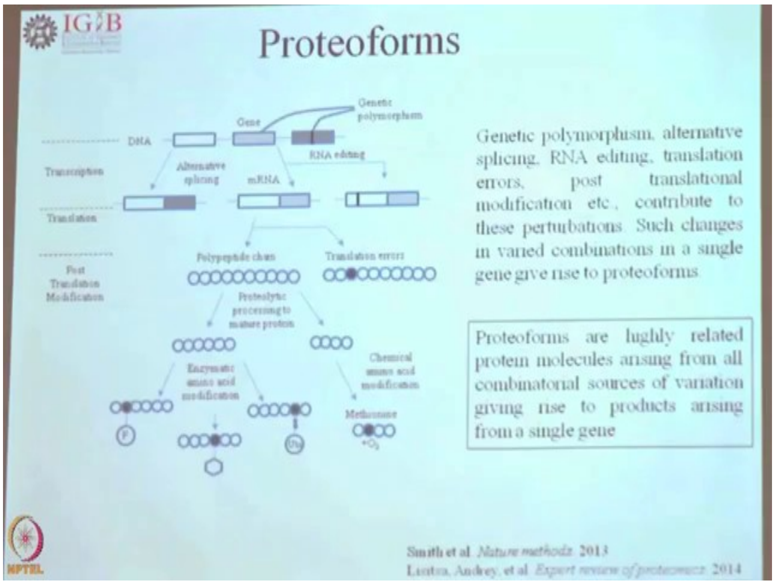
Lecture – 25
Proteomics Data Analysis

Welcome, to MOOC course on Introduction to Proteogenomics. We have finished couple of hands on session about how to use mass spectrometry data analysis using protigy software; you also gone through many of the basic concepts of looking at mass spectrometry data.

Today, we have another distinguished scientist Dr. Debasish Das who is a professor at CSIR, IGIB Institute in Delhi. Dr. Das is going to talk about integrative proteogenomics approaches in understanding of human proteoforms. As you know proteoforms are various modified forms of a protein molecules after different modifications in a living system. In this lecture, Dr. Das will talk about the importance of proteoforms in human system by taking a reference of a research paper by Dr. Ruedi Aebersold and also by sharing some of the work done in his own lab. Dr. Das will also provide you information for some of the repositories available to look at the protein proteoforms. So, let us welcome Dr. Debasish to tell us about integrative approaches, what proteoforms are and their role in clinical biology.

The topic I chose today to share with you is an integrative proteogenomics approach to unravel human proteoforms. So, this title has three terminologies which needs attention: one of course, proteogenomics the conference is on that, the second one is proteoform; we will try to understand what are these proteoforms I am talking about and the third is the integrative approach. And, the integrative approach is what I mostly work on and so, my major thrust will be the approach by which we identify the proteoforms.

(Refer Slide Time: 02:25)



Proteoforms

Genetic polymorphism
Gene
DNA
RNA editing
Transcription
Alternative splicing
mRNA
Translation
Polypeptide chain
Translation errors
Post Translation Modification
Proteolytic processing to mature protein
Enzymatic amino acid modification
Chemical amino acid modification
P
Ub
Methionine
+O2

Genetic polymorphism, alternative splicing, RNA editing, translation errors, post translational modification etc., contribute to these perturbations. Such changes in varied combinations in a single gene give rise to proteoforms

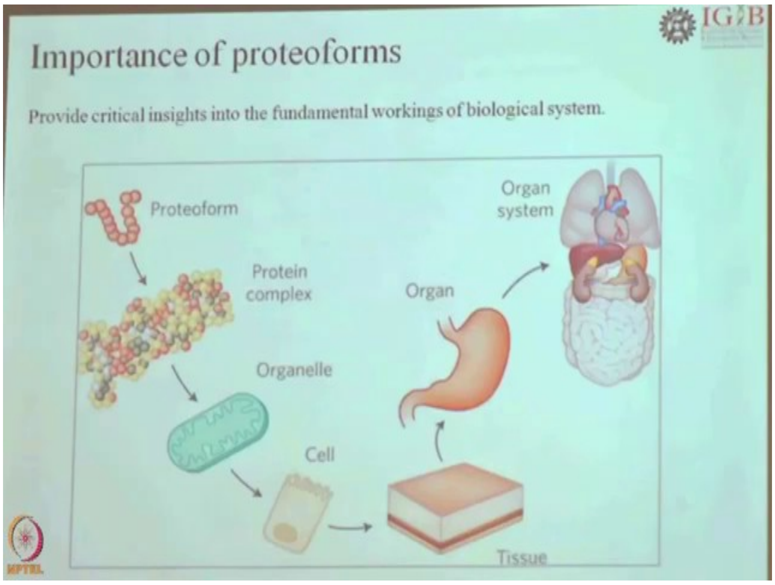
Proteoforms are highly related protein molecules arising from all combinatorial sources of variation giving rise to products arising from a single gene

Smith et al. *Nature methods* 2013
Lisitsa, Andrey, et al. *Expert review of proteomics* 2014

As you can see very nicely covered in Nature methods 2013 and later on in 2014 expert review on proteomics. Proteoform actually the all the alternate forms of the protein which can arise because of alternate splicing, the mRNA and any variation in or the translational errors all of them put together or the even the amino acid modification, all of them put together can lead to generation of a variety of protein proteoforms of a same protein.

Now, all these while we have been talking about the missing protein; so, we had a catalogue of human protein and we were looking for what are the protein that has a transcript evidence and do not have a protein evidence. Now, from there on we move to identify all the proteoforms. So, there are expected to be around 1 lakh proteoforms the number can vary, but this is what people guess and which are those proteoforms that are active, that are functional some of them are involved in diseases. So, discovery of this proteoform actually can give better understanding of the functioning of the tissues.

(Refer Slide Time: 03:35)



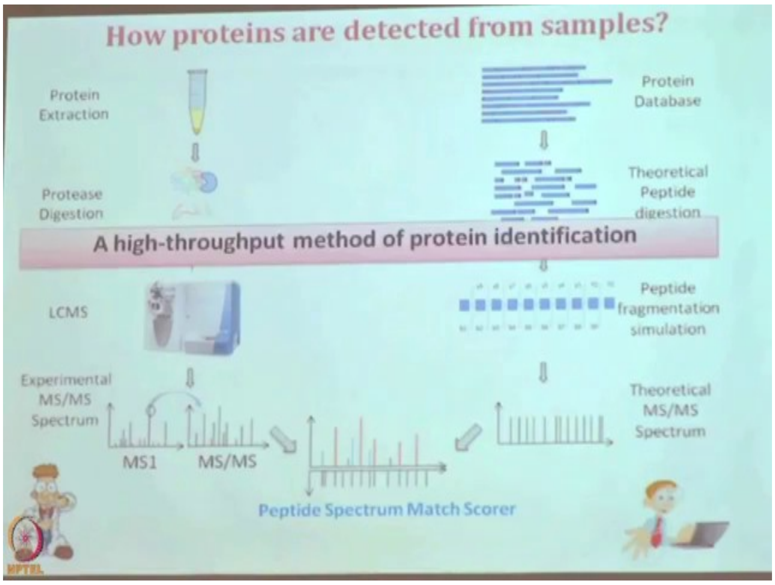
Importance of proteoforms

Provide critical insights into the fundamental workings of biological system.

So, Ruedi very nicely covered this area in nature chemical biology this year 2018 that how the protein proteoforms will be implicated in the future of biology. So, we need to understand first of all where are these proteoforms, where are they expressed and what are the possible roles of these proteoforms we need to understand. But, before we go there we need to understand where are these proteoforms, a tissue why is atlas of the proteoforms need to be done and that is what the research topic on which my student Anurag is in the audience he works on and some of the work that I will be presenting is done by him.

Now, some basic couple of slides have kept for those audience who are new to this area.

(Refer Slide Time: 04:37)

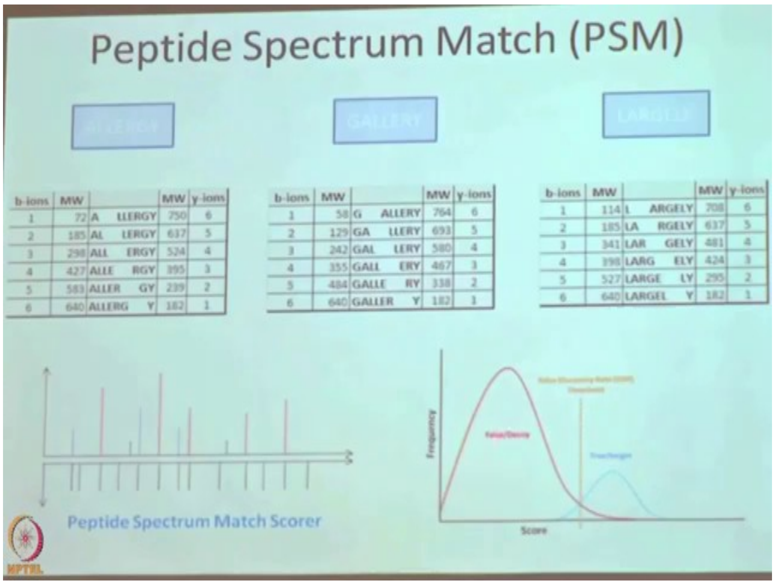


How proteins are detected from samples?
Protein Extraction
Protease Digestion
Protein Database
Theoretical Peptide digestion
A high-throughput method of protein identification
LCMS
Peptide fragmentation simulation
Experimental MS/MS Spectrum
MS1
MS/MS
Theoretical MS/MS Spectrum
Peptide Spectrum Match Scorer

So, how do we detect first of all the peptides in a shotgun proteomics experiment in mass spectrometry? The left hand side which I never do, the experimentalist do. Protein extraction, digestion, injection into the machine and then getting the spectra and my lab starts from here and does the do the right hand side job. So, creation of a database which is very important, unless we create the right kind of database we will not get the answer.

So, this database creation I will little bit delve into this, theoretical digested peptides generation so, this is a rule based, so, there is nothing much here. And, then the peptide simulated fragment generation which will create which will give us theoretical spectra. Now, matching of the theoretical spectra with the experimental spectra and thereby giving a score to this is what is needed.

(Refer Slide Time: 05:31)



Peptide Spectrum Match (PSM)

ALLERGY

b-ions	MW		MW	y-ions
1	72	A LLERGY	750	6
2	185	AL LERGY	637	5
3	298	ALL ERGY	524	4
4	427	ALLE RGY	395	3
5	583	ALLER GY	239	2
6	640	ALLERG Y	182	1

GALLERY

b-ions	MW		MW	y-ions
1	58	G ALLERY	764	6
2	129	GA LLERY	693	5
3	242	GAL LERY	580	4
4	355	GALL ERY	467	3
5	484	GALLE RY	338	2
6	640	GALLER Y	182	1

LARGELY

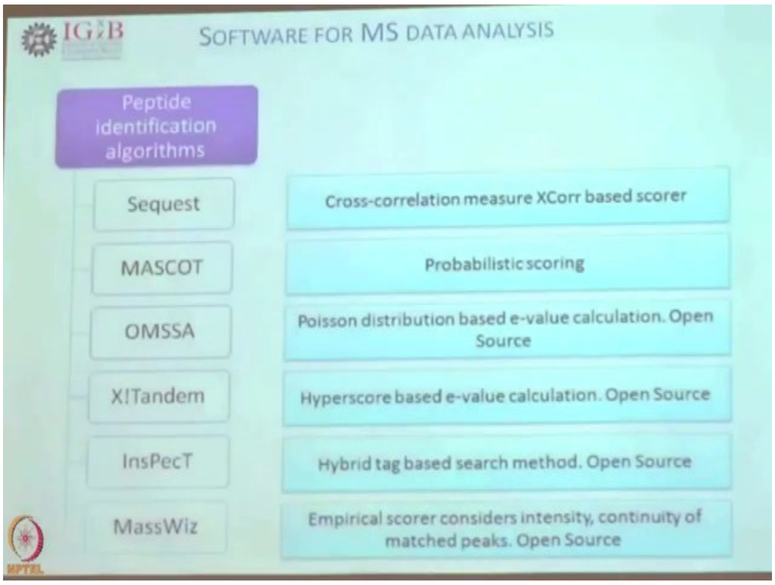
b-ions	MW		MW	y-ions
1	114	L ARGELY	708	6
2	185	LA RGELY	637	5
3	341	LAR GELY	481	4
4	398	LARG ELY	424	3
5	527	LARGE LY	295	2
6	640	LARGEL Y	182	1

Peptide Spectrum Match Scorer

So, here is an example there are three words I have written here allergy, gallery and largely. They are constituting of the same alphabets. So, the amino acid composition are the same, so, but the peptides are different. So, in that case, the answer lies in the MSMS the fragmentation pattern of these the fragments that will generate from these three peptides. And, a matching will be done peptides spectrum match scorer; so, this is different for different algorithms how MASCOT works, how SEQUEST works how tandem works all these scorers will differ in their way of giving a score to this, but however, all of them will get some or the other score.

So, now, it depends on us or on the method to say who has passed and who has not. So, what is the passing score here nobody knows and in fact, that can be a debate here. To do that people take this approach they created a decoy database. A decoy database is a falsified database; database which do not contain the natural proteins. So, the proteins read from right to left maybe or randomized suffering suffered sequences and the target sequence and when you draw a threshold that threshold actually divides the true positives from the false positive.

(Refer Slide Time: 06:57)



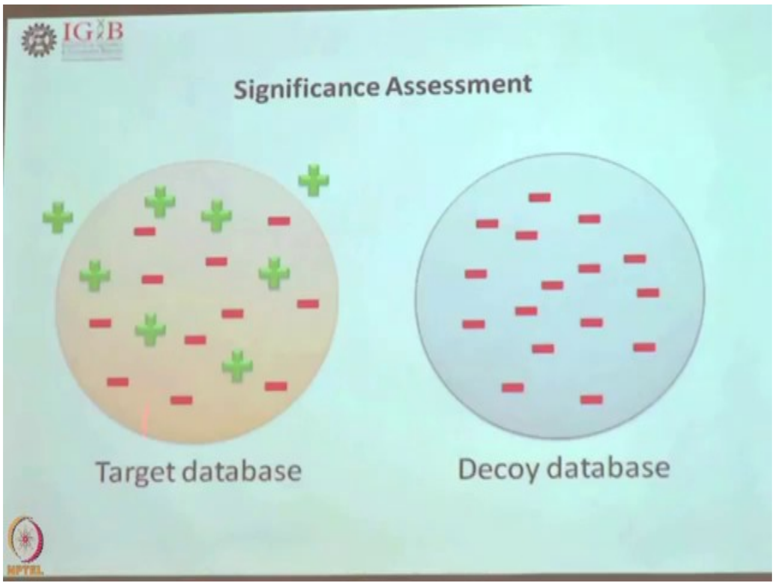
SOFTWARE FOR MS DATA ANALYSIS

Peptide identification algorithms

Sequest	Cross-correlation measure XCorr based scorer
MASCOT	Probabilistic scoring
OMSSA	Poisson distribution based e-value calculation. Open Source
X!Tandem	Hyperscore based e-value calculation. Open Source
InsPecT	Hybrid tag based search method. Open Source
MassWiz	Empirical scorer considers intensity, continuity of matched peaks. Open Source

To do this there are many such search engines are available. Most of them you are familiar; the one that MassWiz is developed in my lab and all others are also available in the domain. And many more also have generally come up come across.

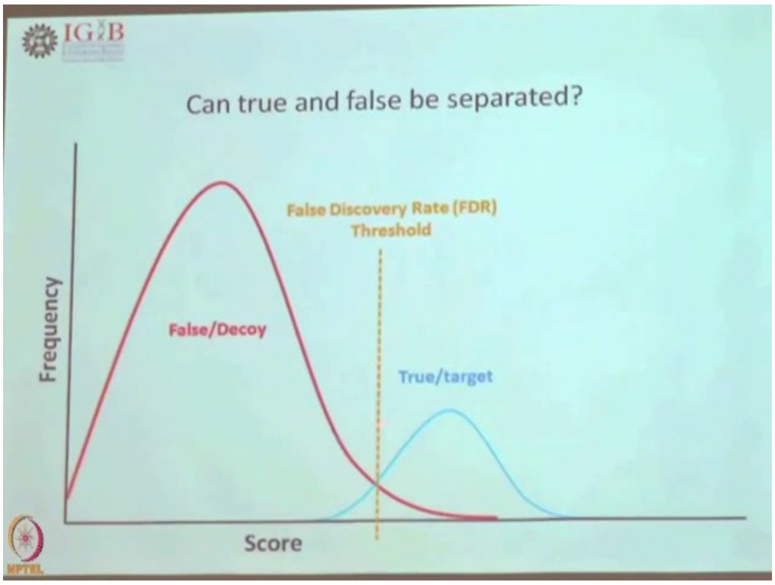
(Refer Slide Time: 07:15)



Significance Assessment

So, what they do is that they give us a lot of peptides that are identified with a score, but then some of them are positive some of them are negative; whereas, in a decoyed database we know for sure that the peptide that we have got and the score distribution now we have got are all false.

(Refer Slide Time: 07:41)

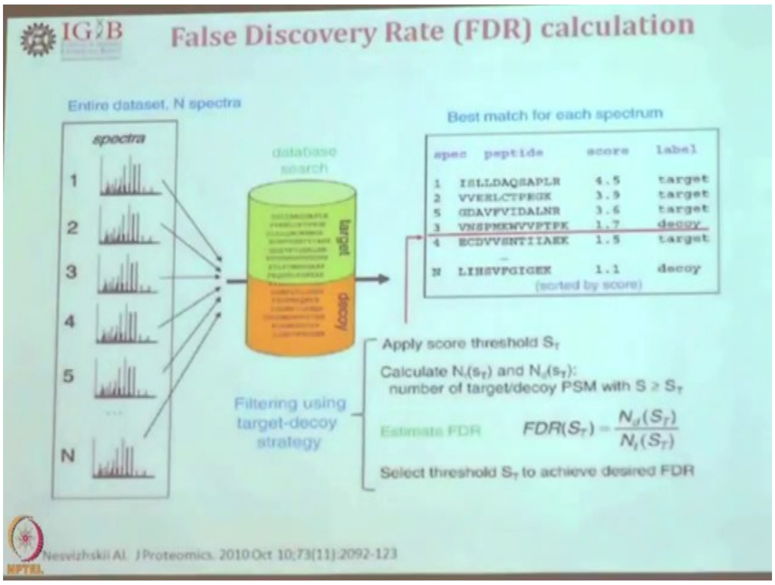


Can true and false be separated?

Now, a comparison between this target and the decoy and a proper threshold will let us know what is exactly the passing score.

So, what is that what is that we generally say the false discovery rate? The false discovery rate is generally calculated like this; every 1 incorrect in 100 correct. So, 1 in 100 is the false discovered rate or 5 percent false discovery 5 incorrect is allowed for 100 true positives.

(Refer Slide Time: 08:07)



False Discovery Rate (FDR) calculation

IGIB
Entire dataset, N spectra
spectra
1
2
3
4
5
...
N
database search
target
decoy
Filtering using target-decoy strategy
NPTEL

Best match for each spectrum

spec	peptide	score	label
1	ISLLDAQSAPLR	4.5	target
2	VVEELCTPEGK	3.9	target
5	GDAVFVIDALNR	3.6	target
3	VNSPMKWVVPTFK	1.7	decoy
4	ECDVVSNTIIAEK	1.5	target
...			
N	LIHSVFGIGEK	1.1	decoy

(sorted by score)

Apply score threshold S_T

Calculate $N_t(s_T)$ and $N_d(s_T)$: number of target/decoy PSM with $S \geq S_T$

Estimate FDR $FDR(S_T) = \frac{N_d(S_T)}{N_t(S_T)}$

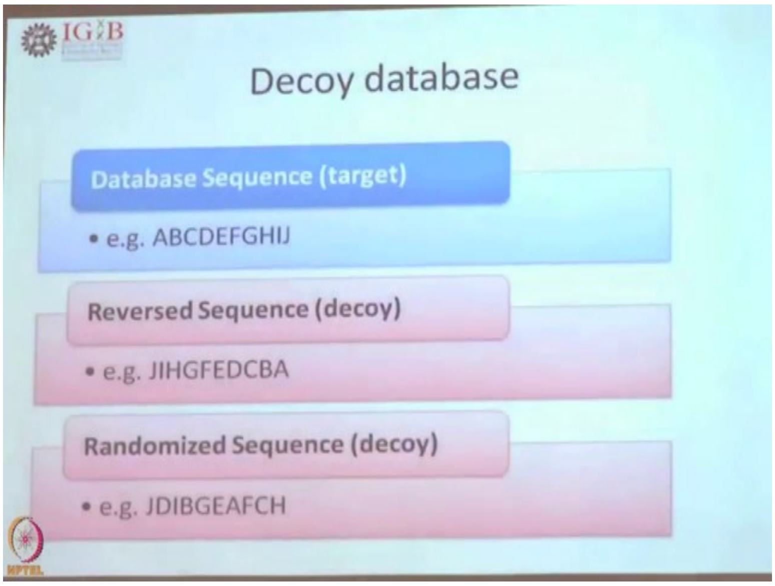
Select threshold S_T to achieve desired FDR

Nesvizhskii AI. J Proteomics. 2010 Oct 10;73(11):2092-123

So, this was again very nicely covered by Nesvizhskii in journal proteomics 2010. You people can go and read this article, very nicely written which says that all the target allocation of the decoy comes. So, FDR is number of decoy divided by a total number of peptides that we have identified from the target.

So, this is how we get to know the FDR at the peptide level at the PSM level. But, then the next challenge will be to identify the proteins FDR protein level FDR. We will see.

(Refer Slide Time: 08:45)



Decoy database

Database Sequence (target)

- e.g. ABCDEFGHIJ

Reversed Sequence (decoy)

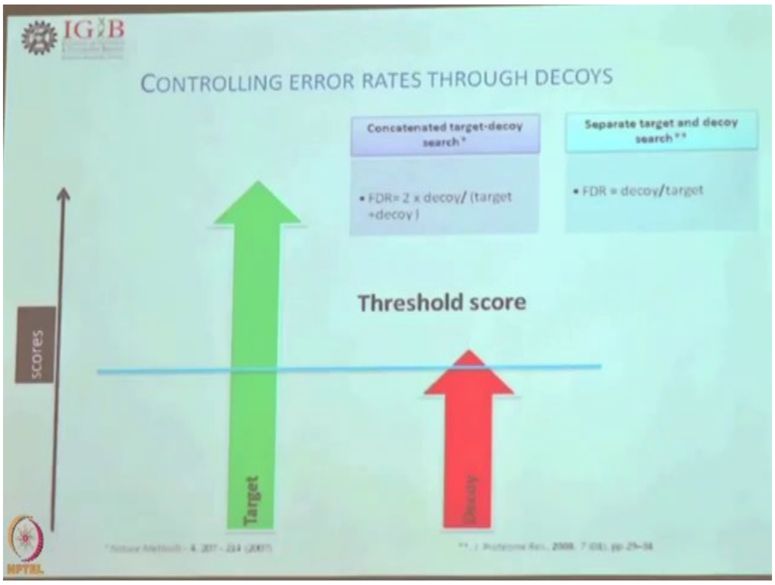
- e.g. JIHGFEDCBA

Randomized Sequence (decoy)

- e.g. JDIBGEAFCH

Now, the decoy database I told is a reverse or randomized sequence alteration of the original sequence so that we can keep the amino acid composition intact, we can keep all other properties of the protein while shuffling all these amino acids.

(Refer Slide Time: 09:07)



CONTROLLING ERROR RATES THROUGH DECOYS

Concatenated target-decoy search*	Separate target and decoy search**
• FDR= 2 x decoy/ (target +decoy)	• FDR = decoy/target

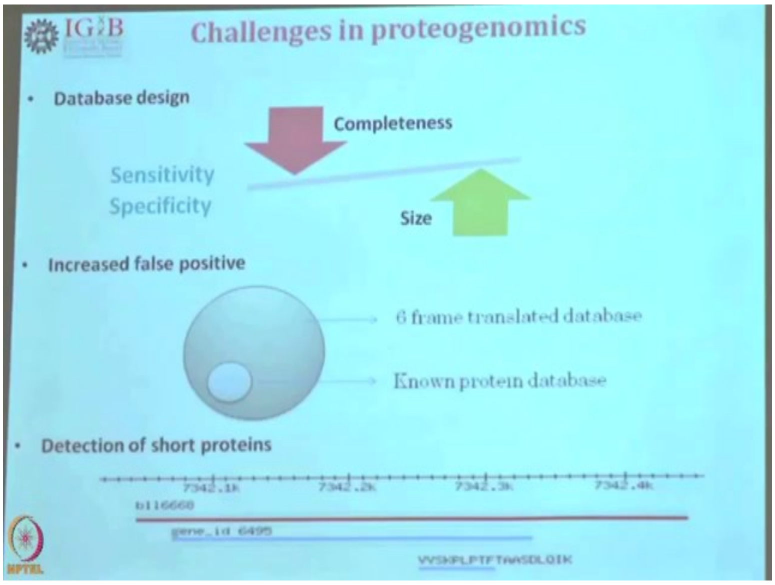
* Nature Methods - 4, 207 - 214 (2007)

** J. Proteome Res., 2008, 7 (01), pp 29–38

And, this is already I have covered. So, this is 0 percent FDR, when no red is above this score no red is there, but this is too purist an approach. So, what we generally do? We reduce our bar in such a way that we accept few allow few rates and get some more greens into our search results and that is how the PSMs are obtained and from there the story begins. We get peptides we match these peptides back to the proteins and from this the protein we infer what are the proteins that are true for our experimental data.

So, there are two ways one can do FDR calculation: one is a concatenated search another is separate. In concatenated what you do actually you merge the target and the decoy into one database; whereas, in a separate search you search in the target separately, you search in the decoy separately and then you apply this formula whatever I just now told FDR is it ratio of decoy to target.

(Refer Slide Time: 10:15)



Challenges in proteogenomics

- Database design
- Increased false positive
- Detection of short proteins

IGIB
Completeness
Sensitivity
Specificity
Size
6 frame translated database
Known protein database
7342.1k
7342.2k
7342.3k
7342.4k
b116668
gene_id 6495
VVSKPLPTFTAASDLQIK
NPTEL

Now, what happens in the case of the whatever just now I said is very generic those who probably did not understand the FDR so, that is how I narrated in a in a brief manner. In proteogenomics case what happens is that you take the genomic sequence you translate computationally and create the protein sequences and thereby you inflate the database size.

So, the database when it is inflated the chance of FDR enormously increases. Earlier we used to search let us say only the known protein from SWISS-PROT we took and search our mass spec data in a limited number of proteins. In case of proteogenomics I take all the theoretical ORFs in case of prokaryotes or translated transcriptomics it in case of eukaryotes and inflate the database size.

Now, when we inflate the database size the chance of false discovery increases. How? supposing you are looking for a place let us say Bhubaneswar and I have given you only the map of Orissa and you are searching Bhubaneswar. So, the chance of once you get it will be correct. I give you a world map and then ask you search Bhubaneswar, but there are chances that by chance you will get another city with a similar name with one letter change here and there and then you start getting confused which is which one is the right one which one is the wrong one.

So, this is the same thing happens as soon as you increase the database size your false discovery rate increases and then you need to do, but you have to do proteogenomics so, database size has to increase. So, you have to find out way how do I limit my false discovery rate even though I search in the larger database, any suggestion? I need to increase my database size because I have to do proteogenomics, but at the same time I want to reduce my false discovery. I want to reduce my false discovery, what do I do? What is the way out? Answer would be there in the next slide, but just for interaction sake. You can be wrong, no problem, but still participate.

Student: Selecting peptides with very high score.

Select the peptides with very high score; that means a Purist model. So, the chances of being wrong will be less that is one way, but you will definitely lose many other correct peptides in the process that is another one way definitely. Any other way?

Student: by increasing replicates

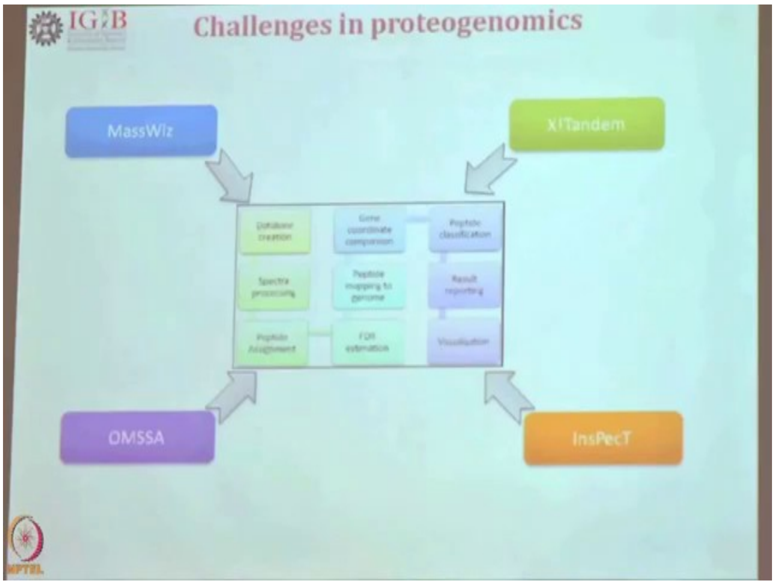
Can.

Student: Replication.

Replication ok. So, two different experiments you look for the same peptide being identified. So, you force me to go to the biologist and ask them to do a replicate study. Of course, that is a good idea always, but we can always increase our search engines. We can use different-different search engines and take their results and hope that multiple search engines will not simultaneously fail in giving you a wrong result.

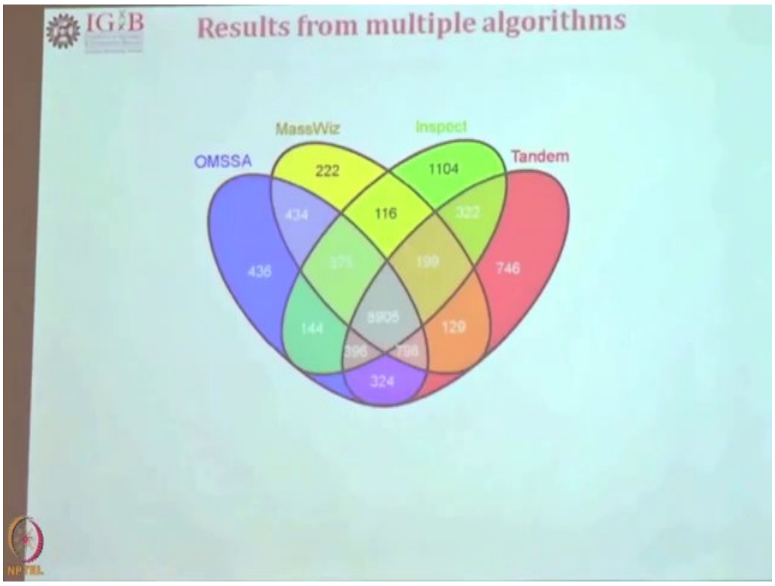
So, this is one way we thought of in the computational lab where can we improve our search result maybe include results from multiple search engines.

(Refer Slide Time: 13:35)


Challenges in proteogenomics

So, what we did we created a pipeline which will do all this process automate automation at the same time take result from various search engine one in house, but others from other sources and take results from all the search engine and then try to analyze the data and hope that the chance of being wrong is less.

(Refer Slide Time: 14:07)


Results from multiple algorithms

What happened when we did this we got a scenario like this, another question coming your way. Now, world is not that simple to me all search engines gave different results. Now, what do I do? Who do whom do I trust? What do you do if you get results like this? same experiment, same database, search parameters being same that search engines are giving you different results. And, you are a PhD student and you have to take a call now, take.

Student: overlapping.

Take the overlapping somebody said consensus from here I think both of both of you are telling the same thing. Any I mean if you are agreeing to this idea no need, but any other any other radical thinking idea is coming?

Student: Elimination of proteins

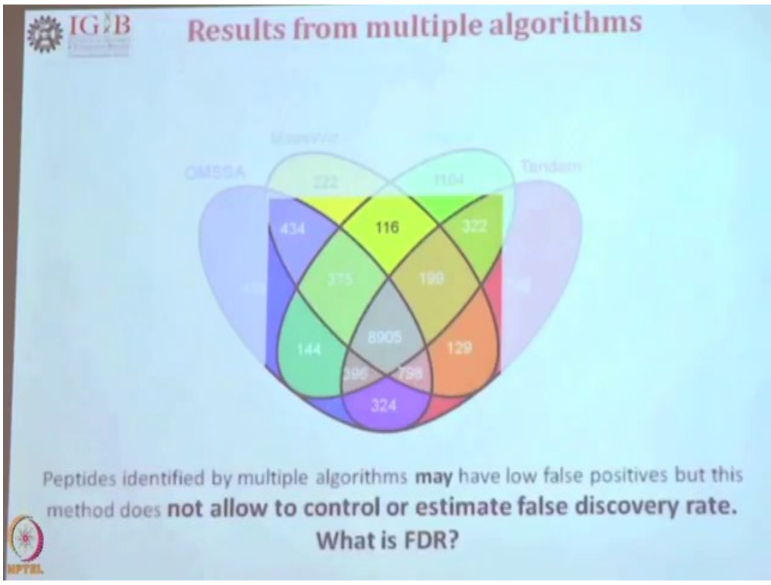
Elimination, on what basis?

Student: On their scores.

On their on their scoring value ok. So, this is one different idea is coming. Compare their scores and then eliminate the weak weaker ones. Now, the problem with me is that a student when coming out of IIT, Bombay, if he gets even 70 percent mark he is smarter student and a student coming out from a unknown university from a remote place he is getting 95 percent, but still is not a smart student. Now, our evaluation processes are not streamline.

So, relying on the score that the student has got was not probably smarter way of doing that. But, of course, we are thinking in that line, can we rescore them? Can we create an entrance examination for all of them to reappear and come through that entrance exams and again? So, something in that line we are thinking, but our first thought was whatever you people you suggested, take the consensus one; easiest probably and little bit safest, but definitely not the smartest.

(Refer Slide Time: 16:15)

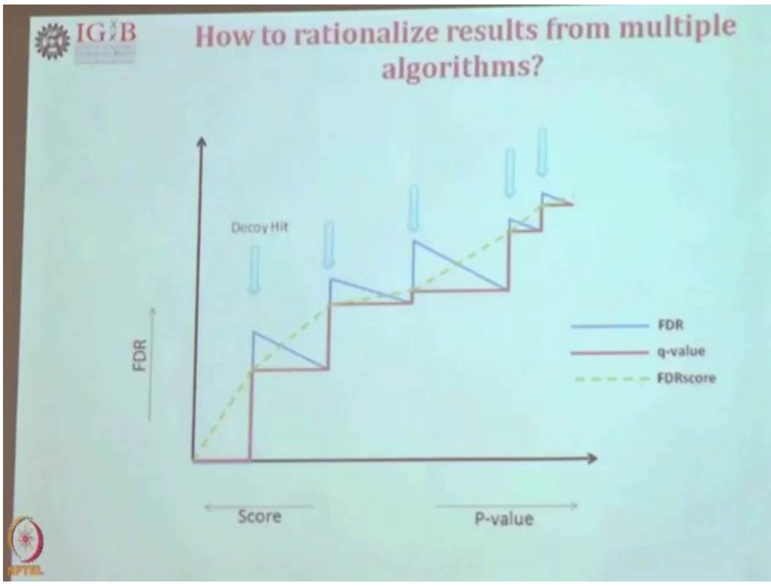


Results from multiple algorithms

Peptides identified by multiple algorithms **may** have low false positives but this method does **not allow to control or estimate false discovery rate. What is FDR?**

So, we went ahead with this, took all those peptides that were identified with two or more algorithms and made our story, went ahead, published and that is how generally we know under pressure you do.

(Refer Slide Time: 16:23)



How to rationalize results from multiple algorithms?

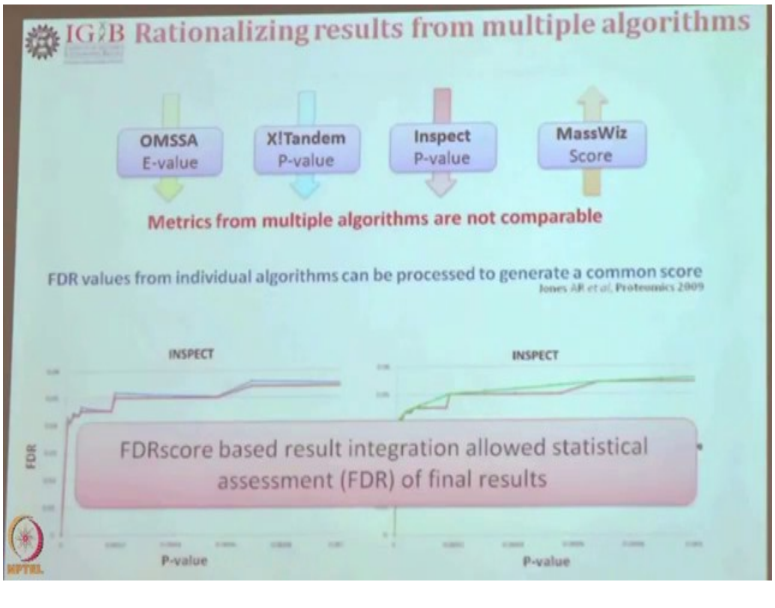
But, then we were not happy with the way I as a computational biologists we handle this problem. So, we started observing what is the behavior of this FDR. So, you look at this curve and try to understand that how the FDR is behaving as the score is reducing.

So, the score is reducing to the right and the FDR axis is on the y-axis and you know the FDR was zero. All of a sudden a red bullet comes the FDR shoots up, and then more and more greens are coming the FDR is going down and then another red hit comes; that means, goes up. So, this is the function by which the FDR is jumping up and down.

Now, this problem with us is that a peptide which is identified at with a higher score had higher FDR than a peptide which is identified with lower score had lower FDR, this is not acceptable. How can you have a person having higher score and still has high false discovery rate. So, what we did? We created a step function and tried to join through a linear line at the base of this next FDR line. And, this was fairly with us because still the FDR is same for this peptide as well as this peptide; for this score as well as this score the FDR has same. But, the best was when we joined these points through a linear regression lines and now, we have a curve which is going upward as the score goes down then you get the FDR is going upwards.

What was interesting is that all the methods irrespective of it is MassWiz, sequest, OMSSA, tandem whatever you take, this green line this behavior of this green line was reserved. So, it was easier for us to create a cut off for FDR score and then use that FDR score for all the methods and choose the peptides.

(Refer Slide Time: 18:45)



IGIB Rationalizing results from multiple algorithms

Metrics from multiple algorithms are not comparable

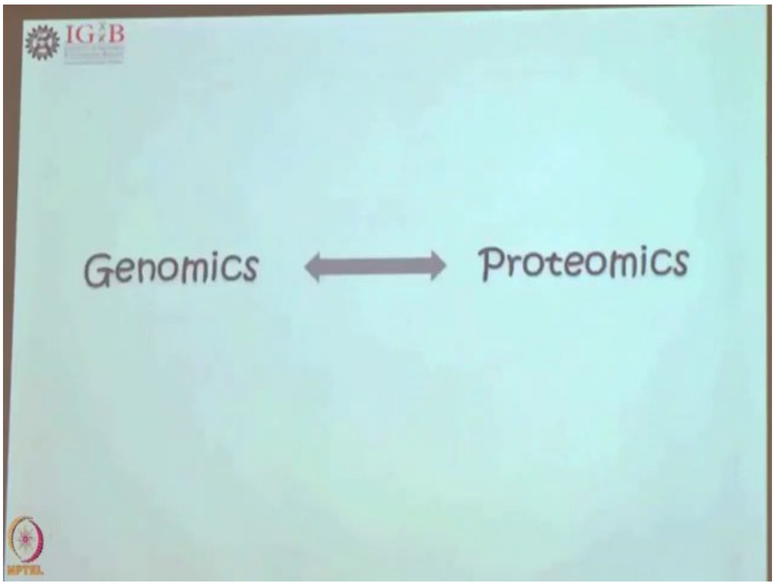
FDR values from individual algorithms can be processed to generate a common score

Jones AR et al, Proteomics 2009

FDRscore based result integration allowed statistical assessment (FDR) of final results

So, what we did, we took the E-value, P-value, P-value score whatever we had the evaluation parameter the metrics we had and then applied on all the methods and at a given cut off for all these algorithms we selected those peptides right which is following that cut off criteria.

(Refer Slide Time: 19:03)

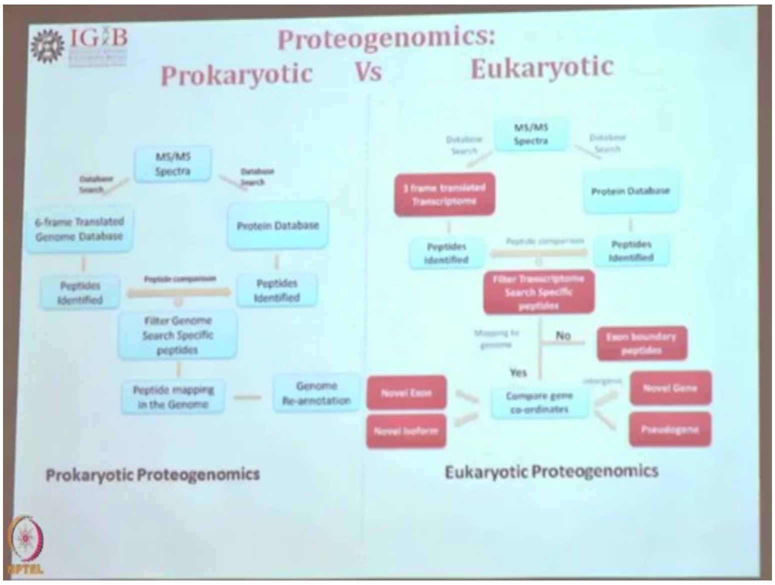


Genomics ⟷ Proteomics

So, that part is over now. Now, multiple algorithms, search results, integration and then getting a pool of peptides from there is what we could have achieve, but the main problem was to identify proteoforms.

So, how do I now get the proteoforms? We have created a translated transcriptomic database, we have now created multiple algorithms and then rationalizing the results from multiple algorithms, now can we have an end to end solution for a mass spectrometry person coming with the data and do a proteogenomics end to end solution?

(Refer Slide Time: 19:51)



Proteogenomics:

Prokaryotic Vs Eukaryotic

MS/MS Spectra

Database Search

6-frame Translated Genome Database

Protein Database

Peptides Identified

Peptide comparison

Peptides Identified

Filter Genome Search Specific peptides

Peptide mapping in the Genome

Genome Re-annotation

Prokaryotic Proteogenomics

MS/MS Spectra

Database Search

3 frame translated Transcriptome

Protein Database

Peptides Identified

Peptide comparison

Peptides Identified

Filter Transcriptome Search Specific peptides

Mapping to genome

No

Exon boundary peptides

Yes

Compare gene co-ordinates

Novel Exon

Novel Isoform

Intergenic

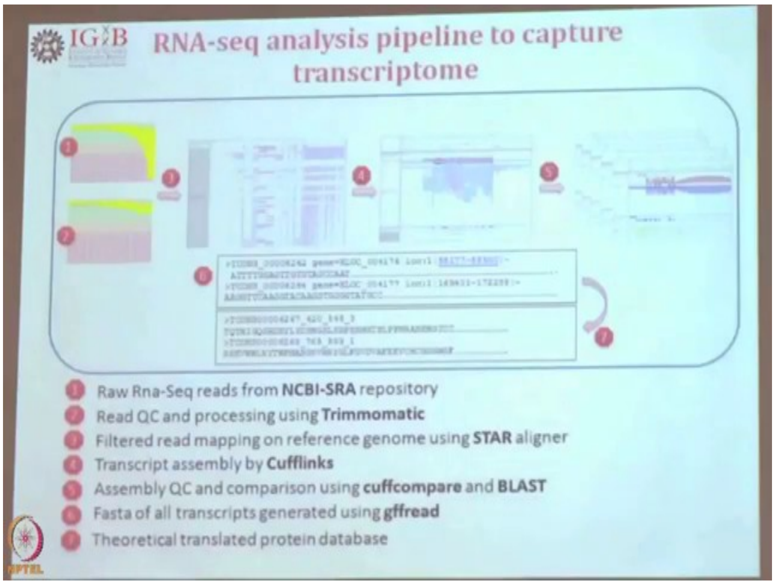
Novel Gene

Pseudogene

Eukaryotic Proteogenomics

So, to do that, we needed a bridge for this, and we constructed this bridge. So, we named it as GenoSuite rest of the talk is will be a little bit boring because that I will beat my own drum, this is what we have done. But, nevertheless just see that what we have done. For prokaryotes was much easier for us; 6-frame translated database creation it was cakewalk and we could get the genome re-annotated with new ORFs identification. But, whereas, for eukaryotes we had lots of difficulty because we had to create the 3-frame print translated transcriptome and then incorporate all the possible alternate splice forms into our proteome.

(Refer Slide Time: 20:33)

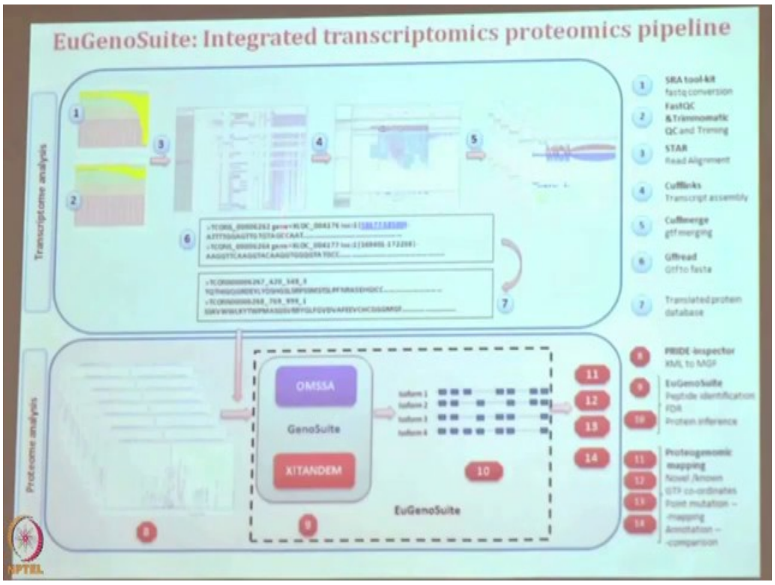


RNA-seq analysis pipeline to capture transcriptome

1. Raw Rna-Seq reads from **NCBI-SRA** repository
2. Read QC and processing using **Trimmomatic**
3. Filtered read mapping on reference genome using **STAR** aligner
4. Transcript assembly by **Cufflinks**
5. Assembly QC and comparison using **cuffcompare** and **BLAST**
6. Fasta of all transcripts generated using **gffread**
7. Theoretical translated protein database

To do that we have created this pipeline by taking the best of tools available elsewhere. So, we did not write any of these codes. So, we took the SRA, Trimmomatic, STAR, Cufflinks whatever was available for analyzing the RNA-seq data, all that we required is a set of protein sequences which represent this transcriptome and this is our subspace.

(Refer Slide Time: 20:59)



EuGenoSuite: Integrated transcriptomics proteomics pipeline

1 SRA tool-kit
fastq conversion

2 FastQC
&Trimmomatic
QC and Triming

3 STAR
Read Alignment

4 Cufflinks
Transcript assembly

5 Cuffmerge
gtf merging

6 Gffread
Gtf to fasta

7 Translated protein
database

8 PRIDE inspector
XML to MGF

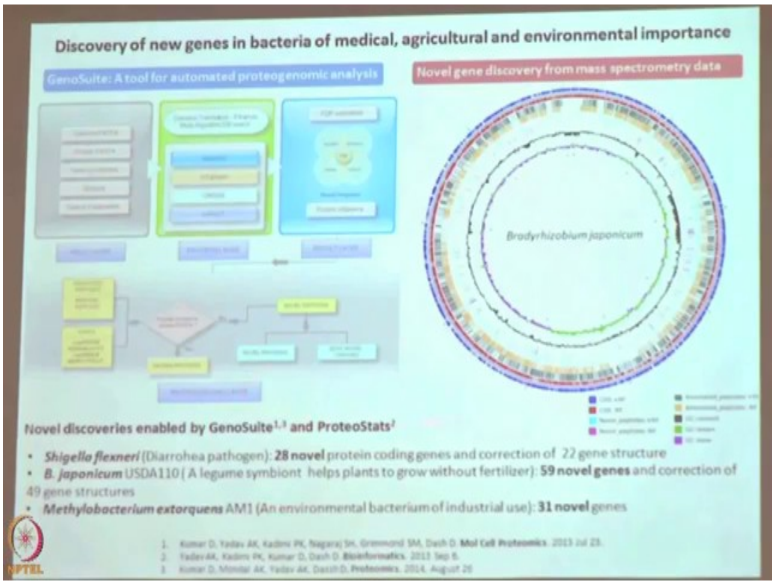
9 EuGenoSuite
Peptide identification
FDR

10 Protein inference

11 12 13 14 Proteogenomic
mapping
Novel /known
GTF co-ordinates
Point mutation –
-mapping
Annotation –
-comparison

And, then using that using multiple search engines we wanted to get the peptides and from peptides infer the protein, I am using the word infer because it is a bottom of approach what we get at the actually the peptides, but what we pose that as if we understand the protein now we know which protein was there. So, from these blocks we infer what are the proteins that we probably would have got.

(Refer Slide Time: 21:17)



Discovery of new genes in bacteria of medical, agricultural and environmental importance

GenoSuite: A tool for automated proteogenomic analysis

Novel gene discovery from mass spectrometry data

Novel discoveries enabled by GenoSuite[1,3] and ProteoStats[2]

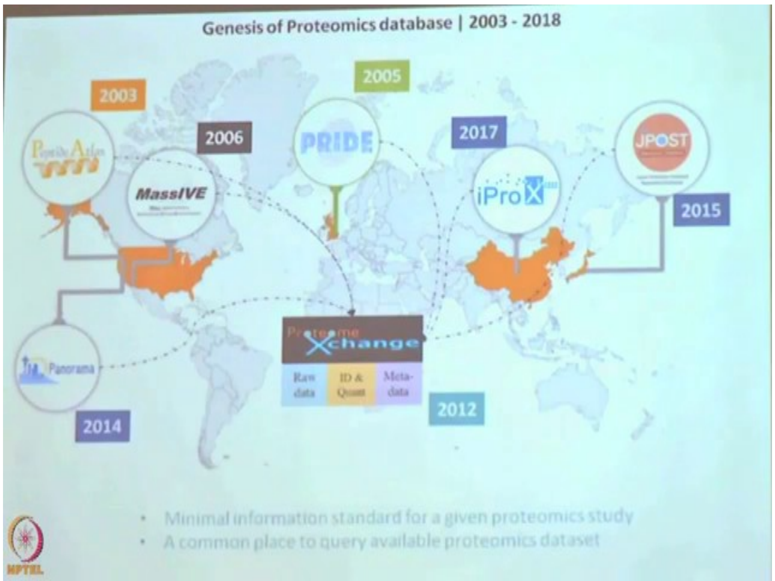
- *Shigella flexneri* (Diarrohea pathogen): **28 novel** protein coding genes and correction of 22 gene structure
- *B. japonicum* USDA110 (A legume symbiont helps plants to grow without fertilizer): **59 novel genes** and correction of 49 gene structures
- *Methylobacterium extorquens* AM1 (An environmental bacterium of industrial use): **31 novel** genes

1. Kumar D, Yadav AK, Kadimi PK, Nagaraj SH, Grimmond SM, Dash D. **Mol Cell Proteomics**. 2013 Jul 23.
2. Yadav AK, Kadimi PK, Kumar D, Dash D. **Bioinformatics**. 2013 Sep 6.
3. Kumar D, Mondal AK, Yadav AK, Dash D. **Proteomics**. 2014, August 26

From the first part the prokaryotic story we published several papers in which used using this particular pipeline. We could identify new translated regions in *Shigella flexneri*, in *Bradyrhizobium japonicum* and *methylobacterium extorquens* and there are if there are students in this audience who are computationally oriented and want to do something. So, here is some low hanging fruit for you as a researcher what you can do; take mass spectrometry data from the internet, take genome proteome data from the internet, use some of these tools and then start re-annotating the genome using the experimentally available mass spec data and the static information of the genome data and you can do wonders by sitting at a in a place with a computer and internet connection you can do all these things.

So, some of these also I could get it done through the trainees who come to my lab and we could re-annotate the genome identify noble translated regions.

(Refer Slide Time: 22:35)

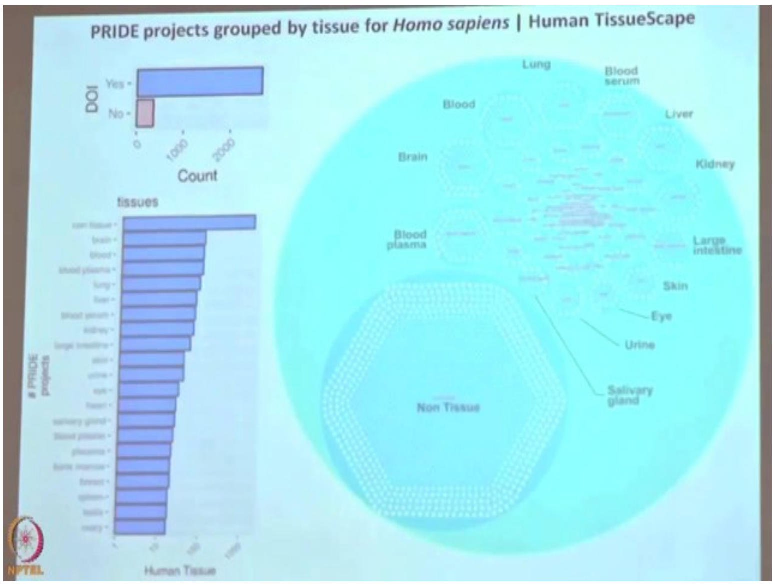


Genesis of Proteomics database | 2003 - 2018

- Minimal information standard for a given proteomics study
- A common place to query available proteomics dataset

So, for that the resource is available already. So, browse the internet you will find many places where you get mass spec data and from there you can download data.

(Refer Slide Time: 22:51)

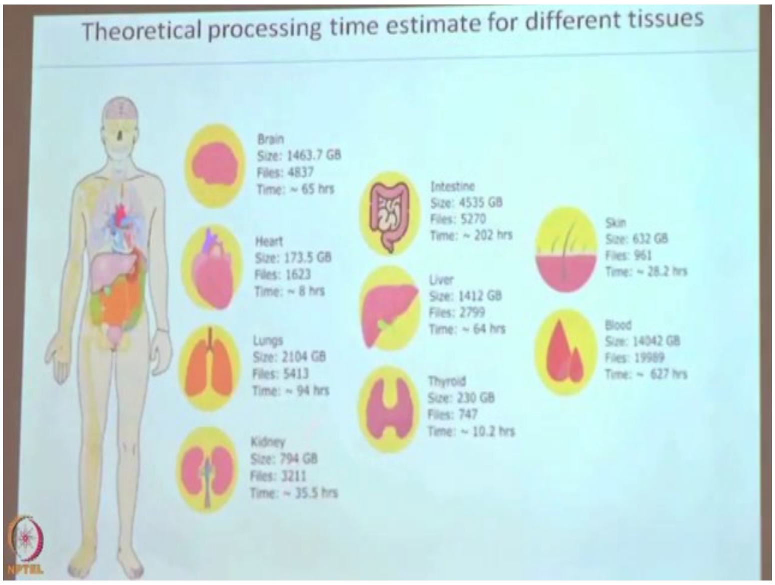


PRIDE projects grouped by tissue for *Homo sapiens* | Human TissueScape

What we have done and for our purpose, since the prokaryotic part is already in some whatever we wanted to do we wanted more a challenging job so, we wanted to go to human proteoforms. So, we looked at the resources and with lots of effort and difficulty we could arrive here, although it looks pretty easy simple to you. From the resource it was difficult for us to identify which are those projects that will give us brain specific, mass spec, blood specific, lung specific and different tissue wise mass spec data because you cannot download the entire data and then re-annotate and then segregate.

We wanted to create a pipeline which will go talk to the pride database massive database and other resources and once you type brain, it will fetch all the brain related mass spec data and give it to you for the analysis. So, for that we created this human tissue scape and this is the statistics of the pride projects where you can see the how many number of projects we have per tissue. And, then we group them on the basis of their group identification DOI or the publication date when it has come and that is how we could group them.

(Refer Slide Time: 24:17)



Theoretical processing time estimate for different tissues

Brain
Size: 1463.7 GB
Files: 4837
Time: ~ 65 hrs

Heart
Size: 173.5 GB
Files: 1623
Time: ~ 8 hrs

Lungs
Size: 2104 GB
Files: 5413
Time: ~ 94 hrs

Kidney
Size: 794 GB
Files: 3211
Time: ~ 35.5 hrs

Intestine
Size: 4535 GB
Files: 5270
Time: ~ 202 hrs

Liver
Size: 1412 GB
Files: 2799
Time: ~ 64 hrs

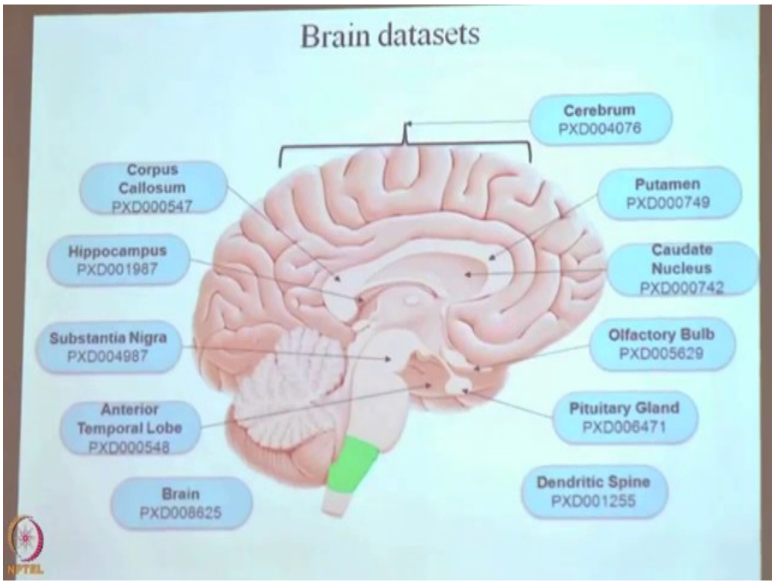
Thyroid
Size: 230 GB
Files: 747
Time: ~ 10.2 hrs

Skin
Size: 632 GB
Files: 961
Time: ~ 28.2 hrs

Blood
Size: 14042 GB
Files: 19989
Time: ~ 627 hrs

As you can see it took about several months for us to analyze a tissue by tissue, what are the where are the proteoforms and after doing the analysis we realized oh this is not the human tissue I was looking for, it is just a cell line. So, a lot of back and forth we had to do. A lot of lesson we had to learn we learned while doing all these things that it is not that straightforward you type lungs and you get it and then you get that the some other cell lines which people have already done analysis.

(Refer Slide Time: 24:51)



Brain datasets

Cerebrum
PXD004076

Corpus Callosum
PXD000547

Putamen
PXD000749

Hippocampus
PXD001987

Caudate Nucleus
PXD000742

Substantia Nigra
PXD004987

Olfactory Bulb
PXD005629

Anterior Temporal Lobe
PXD000548

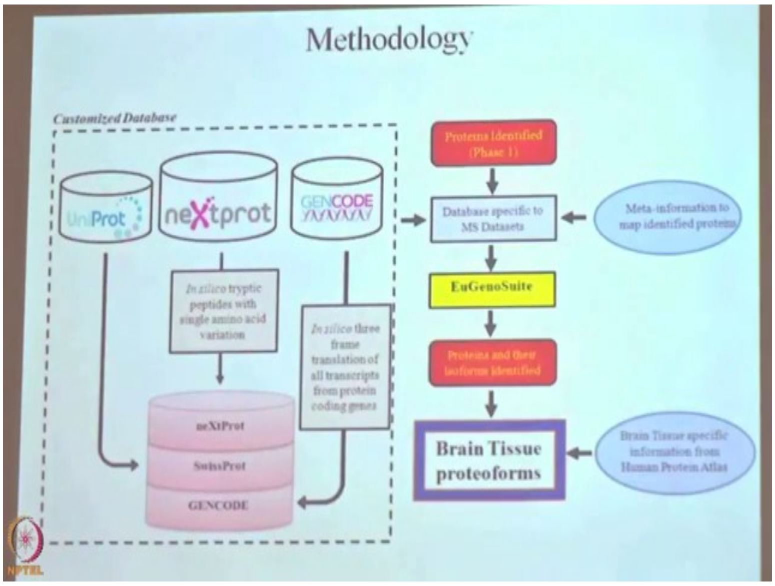
Pituitary Gland
PXD006471

Brain
PXD008625

Dendritic Spine
PXD001255

So, right now my student Anurag who is here he is focusing only on brain and this is only a handful of data sets that we have analyzed.

(Refer Slide Time: 25:01)



Methodology

Customized Database

UniProt

neXtprot

GENCODE

In silico tryptic peptides with single amino acid variation

In silico three frame translation of all transcripts from protein coding genes

neXtProt

SwissProt

GENCODE

Proteins Identified (Phase 1)

Database specific to MS Datasets

Meta-information to map identified proteins

EuGenoSuite

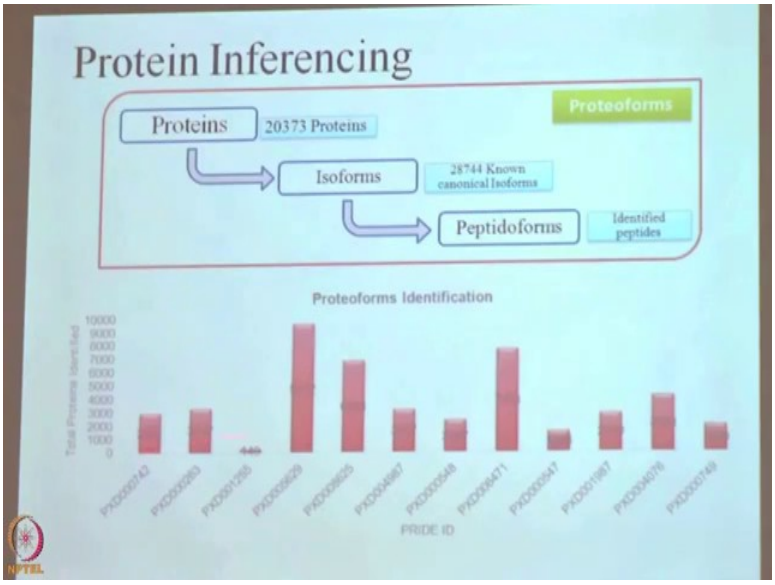
Proteins and their isoforms identified

Brain Tissue proteoforms

Brain Tissue specific information from Human Protein Atlas

Using the this strategy of EuGenoSuite using neXtProt Swissprot and GENCODE.

(Refer Slide Time: 25:09)



Protein Inferencing

Proteins 20373 Proteins

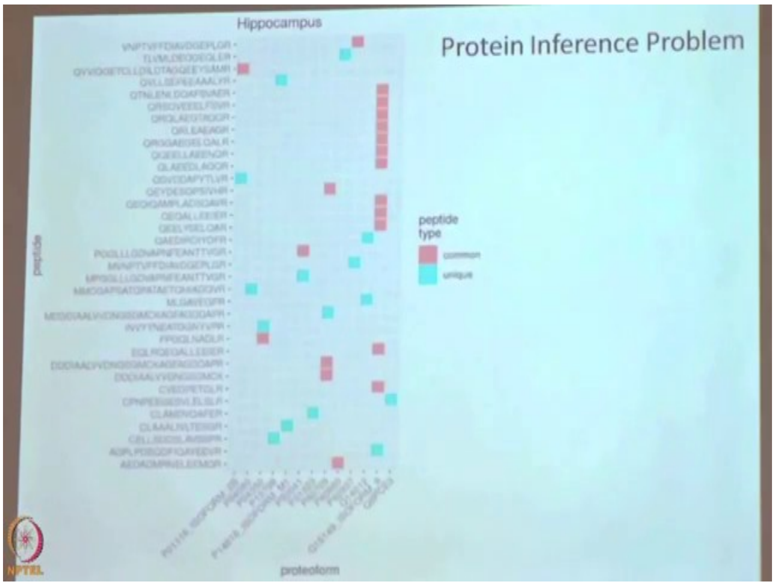
Isoforms 28744 Known canonical Isoforms

Peptidoforms Identified peptides

Proteoforms

We could identify several proteoforms in various tissues and all these proteoforms now, have been ok.

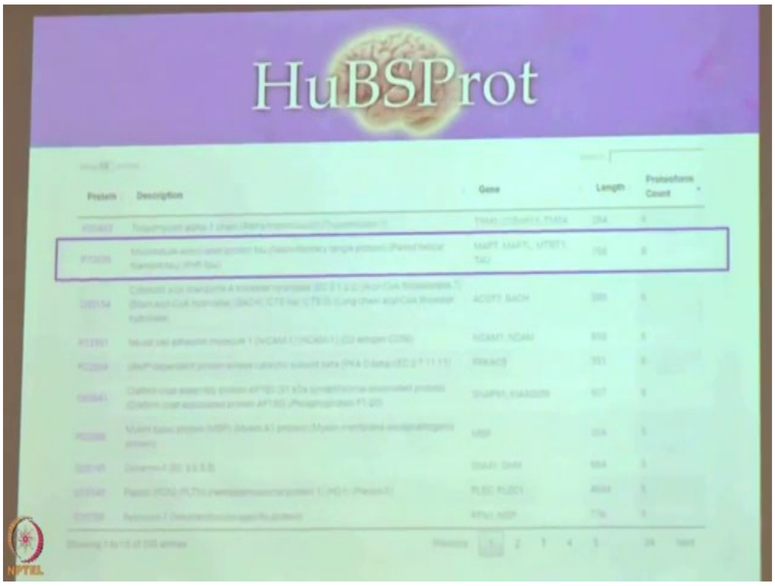
(Refer Slide Time: 25:17)



Protein Inference Problem

This is this is interesting I am I run out of time I need another 5 minutes. Sorry. This is something very interesting another puzzle which is yet to be solved in a computational pipeline manner otherwise right now a lot of involvement is required. See these are the isoforms, these are the peptides. It was much easier for us only when we when we had a unique peptide for that particular proteoform which is not shared with other proteoforms. So, these proteoforms could be identified.

(Refer Slide Time: 25:57)

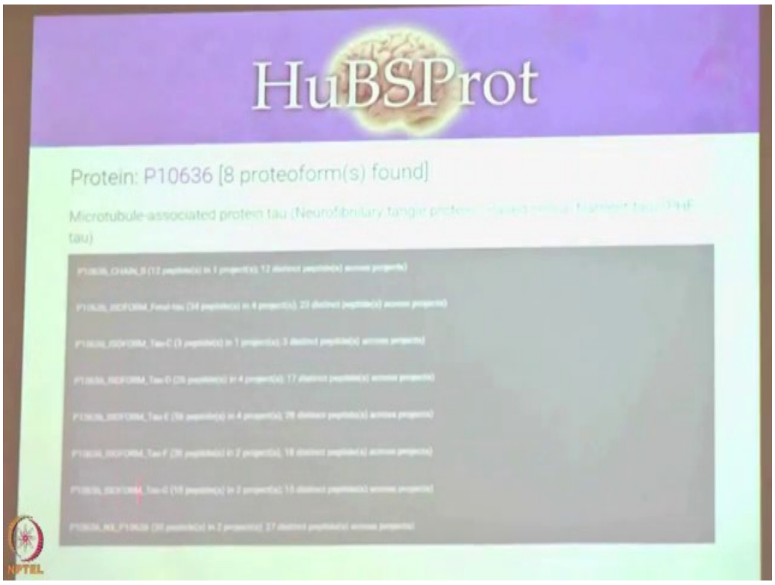


HuBSProt

Protein	Description	Gene	Length	Proteoform Count
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

And, then they have been put into the database. These are all the proteins, their function, their gene names and number of proteoform each of them. For example, tau protein if you look you look into it and you see that these are the various proteoforms of tau available.

(Refer Slide Time: 26:11)



HuBSProt

Protein: P10636 [8 proteoform(s) found]

Microtubule-associated protein tau (Neurofibrillary tangle protein) (Paired helical filament-tau) (PHF-tau)

P10636_CHAIN_0 (12 peptide(s) in 1 project(s); 12 distinct peptide(s) across projects)

P10636_ISOFORM_Fetal-tau (34 peptide(s) in 4 project(s); 23 distinct peptide(s) across projects)

P10636_ISOFORM_Tau-C (3 peptide(s) in 1 project(s); 3 distinct peptide(s) across projects)

P10636_ISOFORM_Tau-D (26 peptide(s) in 4 project(s); 17 distinct peptide(s) across projects)

P10636_ISOFORM_Tau-E (59 peptide(s) in 4 project(s); 29 distinct peptide(s) across projects)

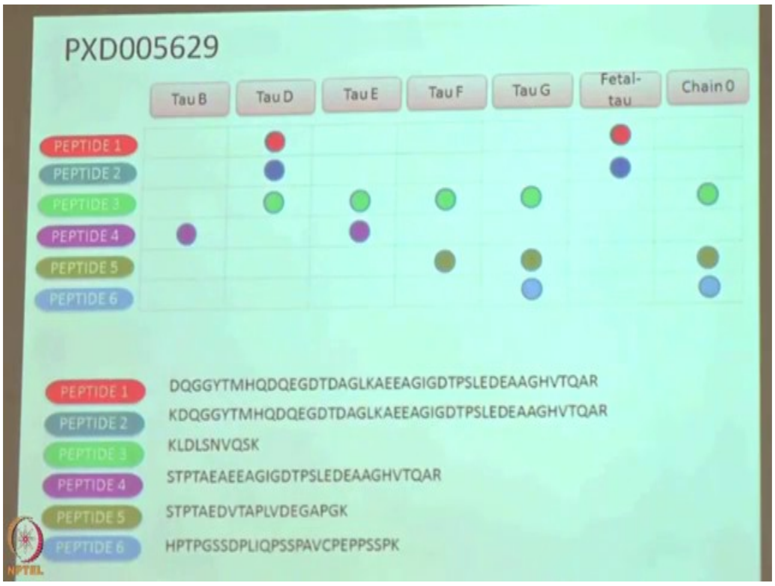
P10636_ISOFORM_Tau-F (28 peptide(s) in 2 project(s); 18 distinct peptide(s) across projects)

P10636_ISOFORM_Tau-G (18 peptide(s) in 2 project(s); 15 distinct peptide(s) across projects)

P10636_XX_P10636 (30 peptide(s) in 2 project(s); 27 distinct peptide(s) across projects)

Seen in how many different projects, how many distinct peptides were identified in that particular protein.

(Refer Slide Time: 26:23)



PXD005629

	Tau B	Tau D	Tau E	Tau F	Tau G	Fetal-tau	Chain 0
PEPTIDE 1		●				●	
PEPTIDE 2		●				●	
PEPTIDE 3		●	●	●	●		●
PEPTIDE 4	●		●				
PEPTIDE 5				●	●		●
PEPTIDE 6					●		●

PEPTIDE 1	DQGGYTMHQDQEGDTDAGLKAEEAGIGDTPSLEDEAAGHVTQAR
PEPTIDE 2	KDQGGYTMHQDQEGDTDAGLKAEEAGIGDTPSLEDEAAGHVTQAR
PEPTIDE 3	KLDLSNVQSK
PEPTIDE 4	STPTAEAEEAGIGDTPSLEDEAAGHVTQAR
PEPTIDE 5	STPTAEDVTAPLVDEGAPGK
PEPTIDE 6	HPTPGSSDPLIQPSSPAVCPEPPSSPK

And, but then this came. What is this? These are different proteoforms of tau these are different peptides and this puzzle is for you not for me. Which proteoform is present if you see a data like this what would be your answer?

Student: peptide 3

Somebody took a stand first. Peptide 3 is that what you are saying? These two?

Student: So, in this case what we have to see there is more about talking of peptides 5 and 6. So, peptide 5 has a 3 proteins.

[FL] The peptide 5 is mapping to three different proteoforms.

Student: Yeah, proteoforms.

And, 6 is mapping to 2.

Student: Yeah. So, 2 we will take but peptide 5 that the tau F, we can say tau F is there.

Tau F is there.

Student: Yes.

Based on these most data you will say tau F is there.

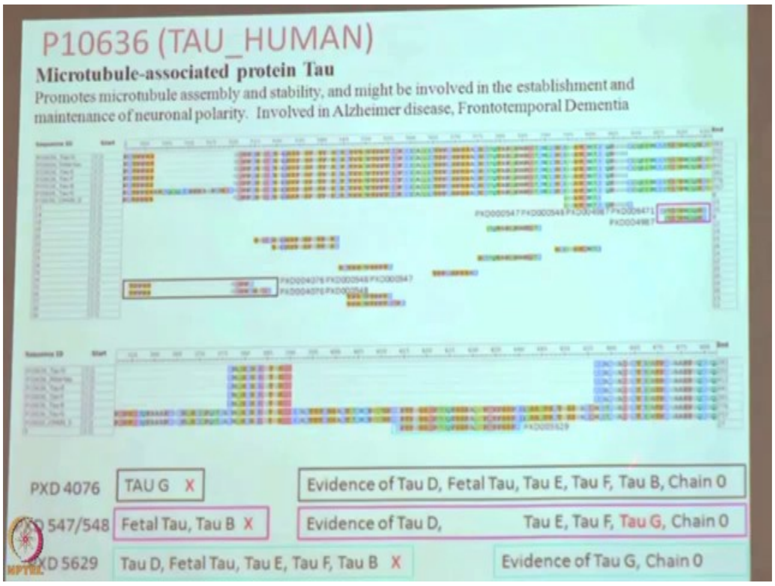
Student: no what are all there

So, there is no simple solution to this problem.

Student: Correct.

Even the answer that I will give you we can spend another half an hour debating on that, but.

(Refer Slide Time: 27:31)



P10636 (TAU_HUMAN)

Microtubule-associated protein Tau

Promotes microtubule assembly and stability, and might be involved in the establishment and maintenance of neuronal polarity. Involved in Alzheimer disease, Frontotemporal Dementia

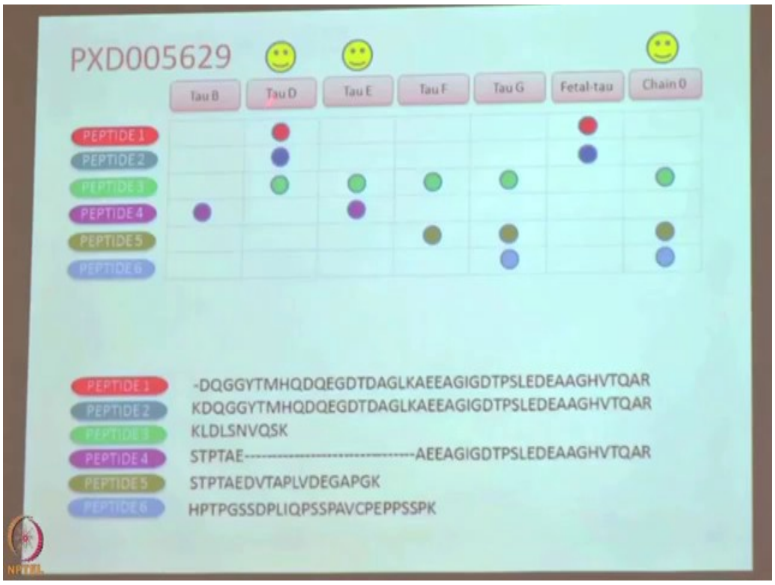
PXD 4076 TAU G X Evidence of Tau D, Fetal Tau, Tau E, Tau F, Tau B, Chain 0

PXD 547/548 Fetal Tau, Tau B X Evidence of Tau D, Tau E, Tau F, Tau G, Chain 0

PXD 5629 Tau D, Fetal Tau, Tau E, Tau F, Tau B X Evidence of Tau G, Chain 0

We chose it like this. We map all these peptides onto their transcript. As you can see for every match we say that evidence for D Tau Fetal E, F, but not G; for the pink peptide that is this one these are all possible not these two possible now these two possible. For the green peptide, where are the green one guy? Yeah, this is here for these green peptide evidence for this, but not this. So, this, but not this; this, but not this, through a series of statements like this and also from other because this could not be shown in one single slide I broke them into another slide.

(Refer Slide Time: 28:17)



PXD005629

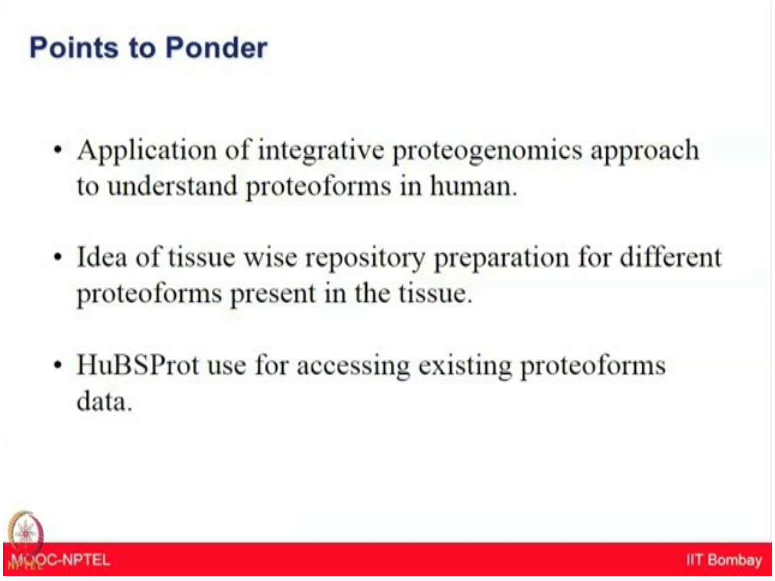
	Tau B	Tau D	Tau E	Tau F	Tau G	Fetal-tau	Chain 0
		☺	☺				☺
PEPTIDE 1		●				●	
PEPTIDE 2		●				●	
PEPTIDE 3		●	●	●	●		●
PEPTIDE 4	●		●				
PEPTIDE 5				●	●		●
PEPTIDE 6					●		●

PEPTIDE 1	-DQGGYTMHQDQEGDTDAGLKAEEAGIGDTPSLEDEAAGHVTQAR
PEPTIDE 2	KDQGGYTMHQDQEGDTDAGLKAEEAGIGDTPSLEDEAAGHVTQAR
PEPTIDE 3	KLDLSNVQSK
PEPTIDE 4	STPTAE-----------------------------AEEAGIGDTPSLEDEAAGHVTQAR
PEPTIDE 5	STPTAEDVTAPLVDEGAPGK
PEPTIDE 6	HPTPGSSDPLIQPSSPAVCPEPPSSPK

So, evidence of Tau D and Fetal Tau comparing all of them together we just given an answer that probably these are the three proteins most likely, I mean the answer is most likely these are the three proteoforms which are there in my sample. But, as I told you very clearly that we can again debate for another one hour overnight on this why this is possible, why that is not possible. So, we have to go back to the data.

And, then you can see look at the peptides the this particular peptide is a unique peptide which clearly says after E the A E comes which is which is very difficult to read from here. So, which will tell that one only one proteoform is possible; other proteoform cannot explain such a separation of the peptides.

(Refer Slide Time: 29:05)

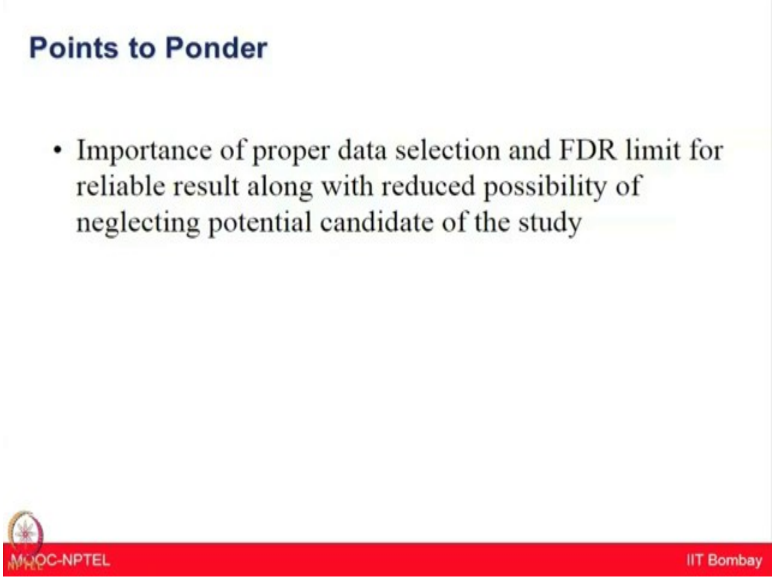


Points to Ponder

- Application of integrative proteogenomics approach to understand proteoforms in human.
- Idea of tissue wise repository preparation for different proteoforms present in the tissue.
- HuBSProt use for accessing existing proteoforms data.

MOOC-NPTEL IIT Bombay

(Refer Slide Time: 29:21)



Points to Ponder

- Importance of proper data selection and FDR limit for reliable result along with reduced possibility of neglecting potential candidate of the study

MOOC-NPTEL IIT Bombay

So, I hope from this lecture of Dr. Debasish Das you got a glimpse of how one can process the proteomics samples and prepare a database to facilitate mass spectra data understanding and analysis. You also learnt about preferable limit and role of false discovery rate in mass spectrometry data analysis. We have also learnt about the hurdles which are related to the multiple algorithms available for data analysis. Dr. Debasish explained about the possible ways to eliminate and how to select the proteins from differently used algorithms.

We have also learned about various sites for database search like UniProt, neXtprot and GENECODE to make customized database for the study. Use of hops prot for accessing already reported proteoforms of a gene could be another valuable resource. So, in the next lectures we are going to shift gears and now Dr. Mani will take you to work flows of automated data processing.

Thank you.