Comprehensive Molecular Diagnostics and Advanced Gene Expression Analysis

Prof. Arindam Ghosh

Dr. B.C. Roy Multi-Speciality Medical Research Centre

Indian Institute of Technology Kharagpur

Lecture 35 : Proteomic Data Analysis and Bioinformatic Tools

Hello students, Namaskar and welcome back to your lecture series on comprehensive molecular diagnostics and advanced gene expression analysis. We are in module 7, where we are discussing high throughput proteomics and today's topic is the final step of high throughput proteomics that is proteomic data analysis and bioinformatic tools. Now we will be covering these concepts basically we will be first looking into how the proteomic data is generated right and then we will be looking into the workflow of how to analyze those data. Then we will be discussing various tools you know the processes or the we divide proteomic data analysis into multiple function number one pre-processing, then identification, characterization as well as quantification of protein, functional annotation, pathway analysis. So we will be discussing tools related to all of them. Alright I will be discussing mainly tools that are available for free and public domain and some cases I might discuss one or two tools that needs to be purchased right.

We will also discussing tools about data visualization and finally we will also be discussing what challenges do we face and how we can improve them and the future directions of bioinformatics in relation to proteomics right. So regarding high throughput proteomics what is high throughput proteomics? You already know that right. So a newly emerging field of life science research that uses high throughput technology. High throughput means ability to deliver maximum number of results from maximum number of samples in unit time per second right per unit time.

To display identify and or characterize all the proteins in a given cell or tissue of an organism that is proteome. So we are characterizing the entire proteome a lot of them in a small time that is high throughput proteomics ok. So the data definitely as is generated from all these high throughput proteomics experiment that we have discussed in these last two weeks in module 6 as well as module 7. They mainly comprise of the identity that is what protein the abundance how much, what modifications how proteins are being changed all right in the given biological sample right. So you already know that the two most common ways or the methods that we use to identify and quantify protein standard are two dimensional gel electrophoresis and quantitative mass spectrometry ok.

So these are the two most used high throughput proteomics processes. And the softwares are generally focused regarding these. If you have to choose one, the one will be mass spectrometry. 2D gel definitely very important very useful, but with the advent of advancement of mass spectrometry we always prefer mass spectrometry. If we are given a choice between 2D gel electrophoresis and mass spectrometry to quantify and separate proteins ok.

So why do we I mean, but there is a challenge or there is a concern all right. The main concern is due to the variability from lab to lab, from preparation to preparation, from person to person all right from equipment to equipment. So variability in sample preparation, instrumentation as well as the multiple range of protein. So since everything is varying there should be I mean there is a problem in ideal representation reproduction of data regarding its accuracy and precision. Of course we are improving day by day.

What we were 10 years back we are leaps and bounds ahead right now, but still there is a challenge. So the aim is since the final step is actually the data analysis the aim being to use a common platform so that everyone can represent their data in a similar way so that all the data from various laboratories can be compared. So we should have a unified vision all right. So the goal is to for everyone to use the common software or common platform or common program. So in this discussion I will be mentioning those all right beyond the scope of this class to elaborately demonstrate the working principle of each and every such software because there are PhD courses, there are postdoctoral courses on this entire thing all right.

So just on one program you can have your PhD course right. So we will be just giving you an overview so that you can note down the software, you can apply them in your own lab, you can explore them at your own will so that you can choose the platform in which you are finding yourself or your work area most comfortable. So previous I mean prior to that the workflow of data analysis involves various steps number one preprocessing all right, then identification, quantification, functional annotation and pathway analysis ok. These are the various workflow of I mean how we can analyze the proteomic data. Now what do we mean by preprocessing? This means cleaning of the data, normalization as well as quality control to ensure reliability of subsequent analysis.

So even before processing we are actually doing something with the data right. So preprocessing definitely a part of data analysis and after that we are doing all these things and there are programs for each and every one of them right and you may choose to specialize in any one of them the field is so vast and so much developing ok. So just by looking at the historical perspective you know field of genomics it definitely boomed with the human genome project right. So it started after 90 and slowly something or the other is always being discovered right. So it is a smear that is going on but the majority the majority technical leap happened in and around mid 90s right.

Compared to that in proteomics the major development of 2D gel was a DIG mass spectrometry happened in the 2010s, 2010s right. Whereas bioinformatic program they started developing they were running hand in hand from right from next generation sequencing to proteomics, but it is being developed. So the most impactful area is still going on we are still to see the best of development of algorithms in bioinformatics. So it is an emerging field all right it is going ahead as we speak. So this lecture might be invalidated in 10 years time or we will have so much to add all right, but in today's date let us see what are the programs of choice that we can use right.

So tools like open MS and proteo wizard ok I will be discussing in brief about each and every one of them. So they assessed in preprocessing steps mainly tasks like spectrum deconvolution. So when everything is mixed it can actually separate the spectra there are programs by which you can untangle the puzzle all right. Retention time alignment you already know what is retention time during liquid chromatography and missing value imputation. So whenever a protein sequence is analyzed there are some false data rate FDR all right which are known as missing values.

So generally missing values lead to loss of sequence loss of information, but now the programs from their concept or from their knowledge of the protein library they are efficient enough they are becoming efficient enough so that they can identify and refill what could have been the missing sequence this is known as missing value imputation. And maxquant and proteome discoverer as you can find by the name proteome discoverer are also commonly used for cleaning normalization imputation to ensure the quality of proteomic data sets ok. So remember these four open MS, proteome wizard, maxquant and proteome discoverer. Now let us look at them in brief. So for every program I will be mentioning the link you can pause the video you can note down the link and you can actually open these websites you can download the program you can fiddle with them some are application based dot GUI graphical user interface based what you can download and install in your system some are scripts that you can run in your own library those who are already into bioinformatics they will know what library I am talking about for example you can use open MS in Python all right Python is a program.

So there are two variations of how bioinformatics is done one is CLI or command line interface in which you need to type code if you are a developer you have to you are developing any algorithm or if you are analyzing you can use the pre-made algorithms in your own native environment for example JupyterScript ok. And another is graphical user interface or GUI means the software you will actually be downloading you will be installing by clicking next and I agree all those commands and then you can simply use there are interfaces where you can do data there are tutorials also available in each of these websites on how to do them. So open office absolutely free you can do proteomics metabolomics there are multiple workflow but mainly it was originally a few method of choice for pre-processing mind it after the end of the discussion you might find that you can use any one of the program to do all of the functions that I mentioned but some are good in one some are better in another all right but most of these programs are being developed in such a way so that they are a one step solution to entire proteomic workflow right. So regarding open MS open source software all right based on C++ library for LCM's data management and analysis all right now it has been modified to work in Python also. So it provides rapid development of mass spectrometry related software so you can actually tweak your mass spectrometry software using this open MS platform again free software runs under any platform many mainstream platform for example Windows Macintosh OS as well as in Linux it comes with the pre variety of I mean variety of pre-built tools all right with each which you can do proteomic and metabolic data analysis for example top tool as well as data visualization.

Remember I will be discussing data visualization software at last but you can use even the first very first software I am mentioning that is open MS for data visualization one-dimensional two-dimensional as well as 3D very beautiful graphs using top view module of open MS right and it also offers analysis and quantization protocol for various processes you have already been taught you have already learned right. Label-free quantification, SILAC, ITRAQ, TMT, Selective Reaction Monitoring, SWATH. I am finding a comfort in teaching all this because all of these programs have been discussed in detail of this process have been discussed in details. So just imagine someone who might have missed the lectures on previous proteomics he or she might be confused in listening to all these terms but you are not right and if you are a bit confused I would suggest after watching this video you can go back and review various lectures on module 7 where you have discussed in details about these processes. Next is Proteo Wizard free software this is the link all right.

So what does put what specialty about Proteo Wizard almost similar the functionality is almost similar they fall in the similar category of software what they can do again open source cross platform software and they can use various software libraries and tools. For example, this MS convert and skyline these are also a sub module of MS open MS right thus they can facilitate proteomic data analysis. So see each and every software was developed in isolation to start with various labs or work groups started developing but then whenever it comes to collaboration one software should be able to talk to each other this is called hand holding and interfacing. So all of these programs are becoming friendlier to each other so it does not matter whether you work in one platform you can always import and analyze and visualize data that have been created in another platform ok this is the main goal. So these libraries help in rapid tool creation providing a robust pluggable environment framework that simplifies and unifies data file access the one I was telling you something which has been developed in other format something which depository might be lying in other environment you can easily import that and develop your own framework create your own experiment run your own set of data analysis in order to finally, visualize them alright.

So it generally performs standard chemistry and LCMS data set computation this is the one in which proteo wizard is known for again free you are free to use this software alright. So we have discussed two open MS and proteo wizard you should at least note down the name so these names should be in your mind even if the links are not right you can always come back and get hold of the links. Next maxquant very very very interesting software so this is the how the website looks like alright we will come back to this website later. So you can I give you a task you can pause just pause this slide or pause what the video is and take a look at every name that you are finding every terms over here right we will refer to this slide later and see how many eagle eye students can relate back to this slide alright. So maxquant is again a quantitative proteomic software package designed for analyzing large scale MS data ok MS data sets which are in large scale it can be detected by maxquant specially if they are high resolution mass spectrometry data alright.

So several techniques are supported including level free as well as labeling technologies alright mind it label free quantification is often abbreviated as LFQ. So if you listen to this terms basically level free quantification again we can quantify that using maxquant alright. Publicly available can be downloaded for free very user friendly there are brochures PDF tutorials there are multiple videos there are webinars that goes on regarding the one webinar that is already going on this regarding how we can use it for level free quantification and there are these are in the archives also if you have missed something you can always learn from the website itself. Next proteome discoverer so these three open MS proteome is a maxquant free software ok. Next proteome discoverer this one is actually developed by thermo fisher so it is often also known as thermo fisher scientific proteome discoverer named after the parent company who have developed it ok.

So this is a unified platform which actually it is actually paid, but free software demonstration is available so you can try and see whether it suits your workflow whether you like the interface then you can definitely download this is the download link it might appear much bigger, but simply you can put proteome discoverer in Google and you will be redirected to the very first link that comes up ok this is the big link, but it is very good it actually helps us to manage data sets for everything. So for statistical analysis biological annotation as well as data interpretation. Identification and quantification of protein of complex protein definitely very very very fluid using various range of proteomic workflows ok and very important the company I mean the developer highlights that this software is specially meant for various protein-protein interaction post translational modification analysis again isobaric mass tagging means i-TRAC label free quantification data independent acquisition data dependent acquisition again everything has been taught we are now discussing the software, but the platforms by which you can analyze them right. So if you are finding confusion in these terms you can again as I said go back and then learn them relearn them revise them and then come back here. So proteome discoverer paid software very good developed by thermofisher scientific available and then can be used for these various methods.

Next we come to identification quantitation see till now I was discussing about actually discussing about pre-processing, but you already have some idea even the software's mentioned before this slide can also be used in identification and quantification right. Anyways so let us see what originally it was I mean so you already know max quant already discussed right. So mascot here since last many many many years it has been 25 years I would say you know almost silver jubilee it is used right mainly for protein identification and these mass quant and skyline are generally used for label free and label quantification that was the choice how they were used, but the boundaries of choices are blurring right now right. So these tools all of these three they help in determining the abundance of protein that is quantification how much protein is there under single dynamic change how proteins interacting how post translational modification is happening how they and that actually helps in identifying potential biomarkers so remember max mascot max quant and skyline. So first we will be discussing mascot very interesting the very first the parental software that is I mean we can say holy grail or gospel of mass spectrometry software the this is the link from matrix science it is available.

Now regarding the mascot it has got multiple domains we are discussing about mascot server alright. So as I told you basically it has been the go-to program the standard go-to program for protein identification from primary sequence databases as well as characterization and quantification by MS. Since mass spectrometry was developed mascot was there almost right. So what does it do it does many things the ones to highlight might be fast and parallel execution combined with probabilistic scoring this is a scoring method again as I told you the mass spectrometry software by analyzing the signal runs a predictive algorithm how it could have been by comparing it with the library proteomic library. So this is known as probabilistic scoring chemical PTM that is post translational modification ITRAQ TMT quantification false data rate alright.

So missing values it can do it top down protein identification so you name it mascot has got everything ok. So there are two modes on how you can how we can you everyone not you only we can use mascot the free service is ideal for evaluation we can evaluate you can run a test experiment on that and then for training smaller datasets and as well as training courses very very very useful you can try out mascot. However if you decide that you need mascot for your lab for routine work for gross clinical sampling then of course you need to license that is you need to purchase the software in your institute level or in your laboratory head ok. For routine large-scale work definitely licensing is required, but remember for small set of experiments small datasets you do not need to pay anything you can use the free web based version alright. Know this we only discussed about mascot server ok which actually when you open the website there are multiple sub domains of mascot.

So multiple work groups are developing multiple programs which are under the single mascot platform. So we discussed about server there is distiller which helps in basically you see distiller I have actually club two screenshots that were strolling down one by one. So what distillers do distiller distillation means you know very I mean purifying impurities from one thing to another so that was the concept right. So basically what distiller does it enables mascot to import data to import the final pure data from various native platforms. There are many native platforms which do mass spectrometry which develop mass spectrometric software for example, Agilent, Bruker those are the leading I mean AB science applied bio science.

So all of them they are developing they are into constant workflow right and mascot distiller actually enables mascot to import their data format run it in a common platform and then helps the researcher to analyze and compare them. Mascot daemon what does it do? It helps in automation it helps in automation and import automation running automation everything you set one program it will do everything on its own. Again mascot parser what does they do? They provide the API keys in order for a developer to develop their own platform using the mascot framework alright. So these are the few technical terminologies which are which might be better understood by the computer science student or the ones who are into developing and understanding bioinformatic program, but for any wet lab researcher mascot server if you are able to navigate and utilize various tools and tips that are already provided in the mascot server it is enough ok.

Skyline very very very important. So this is the interface of skyline again free software we can download it depending on your system whether it is 30 to 64 bit. So basically it is a free available open source Windows client skyline does not work on Macintosh Mac OS right. So you need a Windows based platform or desktop or laptop to operate this thing. So what does it specialize in? It specializes in various I mean mode of reaction in mass spectrometry for example, selected reaction monitoring SRM, MRM alright parallel reaction monitoring data independent acquisition, SWOT data dependent acquisition with must the first MS run quantitative methods. So basically again see we are naming multiple software right which does the same thing, but so that you are actually free to try out everything and use right I am saying the same thing over and over again to emphasize the importance and the completeness of each and every software not only that whenever you are analyzing any mass spectrometry data it is always imperative that you use more than one platform to check and validate you can use free platform suppose you are using open MS and you are using skyline right.

So you can validate the data using two platform and see whether your results are being comparable. So that is basically one way of cross validating the data right. So it does implies cutting edge technologies right. So it is gradually getting refined more and more so that large scale quantitative data can be analyzed with ease analyzed with less time right. This is the story of every software development.

So initially they are developed they remain in immature state we cannot analyze large data large data sets take time, but generally as we go on the algorithms improvement improve and then due to this improvement analysis is much faster throughput increases over time. Next we will be discussing about functional annotation tool. So functional annotation this means assigning biological functions to identified proteins. So now that you have identified proteins so we can annotate basically label their function. So this actually helps to interpret large scale proteomic data sets and they are tools by which function annotation can be done.

Two most important are David and Panther very interesting names right. So basically there are examples of tools that categorized proteins based on gene ontology terms. Ontology is actually the science of existence ok. So they basically they can label the genes based on the proteins and thus helps us to enable or to gain functional insights right. So first discuss David maintained by NIH that National Institute of Health America right.

So the name David actually is an acronym that is derived from the database for annotation visualization and integrated discovery. So what does it do it provides a comprehensive set of annotation tools for investigators to understand biological meaning behind large list of genes. You see whenever we are discussing proteomics you saw the genomics is being developed into proteomics and then it is being converted to I mean it is taken up by bioinformatics. So everything is running in parallel so the functional annotation softwares they actually basically annotate genes which helps us to understand proteins. So very important in proteomic workflow but technically they are gene annotators right basically mRNA.

So these tools are actually based on the David knowledge base which is again built up on the David gene concepts which together pulls up multiple sources from various literatures and helps us to enable these functional annotations. So what does David do for any given gene list David are able to do many things but I have enlisted few things that it can do with respect to proteomics. For example it can list interacting proteins, it can explore the names of those genes that are coding for these proteins in batch, it can link gene disease association, it can highlight various protein functional domains and motifs in those genes, it can again based on the information it can direct them based on what literature the software is doing. So many things are done and lastly it also helps us to visualize these genes in the platforms like bio cut and keg will be again discussing these platforms as well and many more. So this was about in short about David all right if you think the class is overwhelming I am going too fast pause reread again come back ok.

Panther what is this basically again an acronym so the it is has been developed it is this is the website in which the database is there so protein analysis through evolutionary relationships ok classification system. The classification system was designed to classify proteins and their genes first very basic in order to facilitate high throughput analysis right. So basically it is a the core content is basically a gene library annotated gene library right and this genes all these genes the proteins coding for they are can be annotated you I mean to develop phylogenic trees all right. So you can see all nodes in the tree have persistent identifiers that are maintained between various versions of pan-fascin they have been updated, but specifically the one who worked with this older version of software if they come to newer version of software they will find their existing phylogenetic trees their markers still intact while new markers new proteins are being discovered so they are being now added all right. So this actually ultimately provide a stable substrate for annotation of protein properties like sub family what proteins sub family do they belong to what is their function so any new proteins can easily well fit if you do not know just enter that in sequence in pan-tharand they will automatically tell you what can be the functionality and what sub family do the protein belong to very important regarding protein discovery and functional annotation this is how it is look it looks like so you can simply enter the ID for example, either you can upload your own protein you can download in software sequence from to try out from uniprot.

com ok. This is a very big database uniprot uniprot that has got a big database which has got all the proteins and listed discovered till now and still being discovered you can simply paste one in this of in this window which will easily find if you open this web and they will very beautifully give a pictorial diagram where the protein functionally belong to and what family sub family it is what is function so on and so forth. Next pathway analysis pathway analysis what does it do it helps to elucidate biological significance of identified proteins by replacing them within the context of biological pathways of what proteins are play a role in what biological pathway that this process gives us an idea about. So, regarding pathway analysis again basically there visualization platforms of what does it do not strictly visualization platform they have got multiple basically pathway analysis gives us a comprehensive idea about the protein function much like functional annotation but also where do they fit in the physiological pathway. So, the platform that we prefer are biocarta reactome and kegg kegg again a acronym will be discussing very soon. So, what does it they do they reveal the potential connection and functional implication of protein right.

So, how protein are interacting protein-protein interaction very important can be visualized using these platforms to start with biocarta this is again the link alright you can see this has got only so many views. So, attribute pathway you can actually select what pathways you are interested in that way you can choose a protein to work with or if you discover some protein you can again tally that from the biocarta database alright. So, basically it is a community fed database. So, everyone from the communities contributing into the development of this database. So, what does it feature they feature extensive collection of maps they are directing common metabolic pathways signal transduction pathway and many other biochemical pathways alright.

So, what protein are fitting into what pathways and how they are interacting dynamically alright the dynamic maps you allow users to observe how genes interact very very very interesting ok. So, they actually have got more than monolux. So, 120,000 genes from multiple species ok. Now, this is basically the feature set of biocarta which is simply have been pasted from their website ok. So, just see what is the span of information that you can get pathway mapping metabolic pathway biosynthetic pathway signal transduction protein biosynthesis RNA protein biosynthesis cellular processes target best structure classification skeleton based hormonal signal transduction pathways cell growth apoptosis receptor everything how you can imagine any biological process any immunological processes and the information about all those proteins are already there ok.

So, it is the ocean for you to explore one very similar is reactome again pathway browser analysis tool simply find reactions proteins and pathways simply put the reaction you might not be discovered you might be interested into know a function of a protein unknown protein which you have found out in a literature it might not be mentioned clearly just put the name of the protein or name of the gene it will show a very clear pictorial representation will be there where you can easily visualize. So, I would suggest after this is done the lecture is over you just go to each and every website and explore right it is very very very interesting, but again for the sake of limited time it is not possible for in this course in this lecture to demonstrate everything maybe we will come up with the bioinformatics lecture course that will also help us to further understand and find you into inculcating you invoking you the interest in order to explore all of them by a gentle hand holding manner. KEGG the next program is KEGG Kyoto encyclopedia of genes and genomes. So, many things are already available when it started naturally nothing I mean not all I mean features were available. Now there are multiple links which does multiple function alright for example, I am discussing KEGG in relation to pathway database.

So, just take any protein you so these are the sub classification or sub category based on which you can select a pathway you can assign very beautiful colors to those genes and ultimately you will be finding this type of diagram. So, how one pathway leads to another how one gene controls one enzyme how one gene controls one protein everything will be given there is a coloring GUI in which you can design your own maps there is a color there are multiple color specification. There are many things you can do using KEGG. Moving on we are almost at the end of our journey.

So, data visualization. So, whatever data we are analyzing it is always good if we have got a good pictorial representation. Now all of these software that I discussed till now have got some aspect by which they can reproduce their data. So, maximum for quantification will be information there will be numerical numbers, but this data visualization platform this specifically I mean try to represent the proteomic complex data interpret the and communicate the complex data in a good pictorial way. So, that they can be easily visualized easily perceived by any observer alright. So, what are the tools mainly I will be discussing two Perseus and Cytoscape ok.

So, they facilitate visualization of protein-protein interaction and network analysis data. So, see Perseus see this is the figure that have been I have taken from their website only. So, they do multiple things they do classification, they do high intensity plot, they can again annotate the gene and import them in KEGG, they can plot interaction between DNA protein, protein-protein interaction, post translational modification, graphical representation of impute I mean imputed data. So, these functional gene loss they can collect them, network analysis, density plot anything again matrix distribution of protein. So, the interpretation and the way you can represent are manifold very beautiful diagrams can be generated using them.

Now, for the eagle eye students who already does remember that we always already mentioned the Perseus or Perseus term was already mentioned in a previous slide.