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 60 : Integration of Multiomics Data in Molecular Diagnostics

Hello everyone. Finally welcome back to our last lecture series on comprehensive molecular diagnostics and advanced gene expression analysis. And our today's topic is integration of multi omics data in molecular diagnostics. And we will be covering that topic under these following headings we will be first discussing what is omics data and scope of omics data, alright. We will be recapping what are the various types of omics we have studied till date and if we have not touched any topic I will be telling you what all you need to know in order to integrate all the omics data will be this is what are the various models how we can integrate this multi omics data and what are its benefits and challenges and what are the future directions in this omics technology, right. So, omics technology as you already have understood is basically we can addition of I mean whatever there is any bunch of science, right.

Whenever we add the word omics, right it now becomes something comprehensive or global, right. So, a set of molecules if we add omics it becomes a global assessment of the set of molecules, right. So, each type of omics data and there is a huge amount of data that is generated from each type of omics technology. So, each of omics data on its own typically provides a list of differences associated with the disease.

So, if we concern the disease as an end goal our final target is the health for all the precision medicine, right, personalized medicine. So, if we concern any disease and then look up all the possible set of molecules that lead to production of this disease or lead to development of this disease we can study various genes, proteins, metabolic parameters. So, each I mean each and every omics gives us a huge set of data, right. And this omics technology is largely driven by technological advances that have made it possible for cost efficient and high throughput analysis of biology molecules that we have studied extensively in genomics and proteomics in our model, right. Now, there are two concepts of medicine that we briefly discussed in pharmacogenomics and precision medicine.

The traditional medicine which is based on signs and symptoms, right, it deals with symptoms and then treating the symptoms, right. Whereas, the this is transitioning into omics, right, deals with how to prevent that. So, we need to know everything about the disease each and every causative factor at every molecular level so that we can prevent the disease from happening all together by early intervention so that we can improve the quality of life. So, traditional approach has been the patient presence with something and we are treating only after it has developed. However, the systems biology, the omics technology deals with knowing everything, what is the need of omics technology, it is knowing everything so that number one we can prevent the disease and we can precisely prescribe I mean an individual or a patient that medicine which will be needed or applicable for him or her only without any side effects, right.

So, what are the scope of omics data? The scope of omics data I mean what can we do with so many data? Number one very important markers of disease process at various levels at the genetic level, at proteomic level, at microbiome level, at every level we can set and identify new markers based on which even before the disease has happened or this is the start this is pathology starting to happen we can identify these markers and very important these markers are the road for early diagnosis, right. So, differentiate between biological pathways or processes between disease and control groups. Not only that since we are also analyzing experimental data we have got in detail multiple I mean multi-omic data where we can functionally annotate an entire pathway to understand what is going on, what is going on in our experimental system, what is going on in our therapeutic model, right. So, multiple control groups therapeutic groups can be easily compared again at every level with the help of this multi-omics technology. Very important it gives us a greater understanding of the flow of information by studying isolatedly studying genetics which may not transform one is to one into transcriptomics which may not transform one is to one into proteomics, but if we are integrating the whole thing it will help us much better to easily understand what is the flow the entire central dogma in application will I mean it is very it will be easier for us to understand, right.

And this is the cornerstone of precision medicine as I have already discussed in pharmacogenomics module the end goal is either one medicine for me that will be only the most effective for my I do not need to take a general medicine which is there for everyone, right. So, all of this can be achieved with the help of multi-omics data. So, this is the huge scope of multi-omics data which is still in the process of development. Now, recapping what are the various type of omics that we are studying number one genomics, right. I do not need to tell you in much details what is genomics, right because this is the most mature field of omics.

This is the first thing after human genome project entire field of omics revolving around gene has sprung up and multiple genomic test diagnosis laboratories experiment as well as many research and development are going on. It mainly focuses on identifying genetic variants associated with the disease. Everything we are concerned about disease. So, genetic variation why the disease is happening and finding various genetic variants. So, either it can be single nucleotide polymorphism, it can be any mutation whatever any molecular diagnostics related to gene level comes under the purview of genomics, right.

So, whether it is associated with any differential response to treatment or differential response to any outcome of any drug therapy or any side effect or even if related to future prognostics. If the answer is in gene, if you are searching any answer in gene, it falls under genomics. Epigenomics I hope you already know that, right, but still I am trying to recap for those who might have not brushed their previous modules well, well I always assume you are well learned till this point. So, it is always advisable whenever you find something interrelated you just revise the previous lesson, then come back to watch this lecture series. Anyways, epigenomics.

So, you know epigenetics is actually the study of reversible modification of DNA, right or DNA associated protein. Suppose DNA methylation, histone modification, histone acetylation. So, all these things if it is related to genome wide characterization, it is known as epigenomics. Transcriptomics, the name itself signifies it is again a global comprehensive study of all the RNA transcripts that are present. Now, this transcriptomics can be qualitative or quantitative.

So, what is qualitative transcriptomics? Qualitative transcriptomics deals with what are present, right. So, which transcripts are present, identification of any novel transcripts related to a disease, any process of RNN, editing, discovery of any alternative novel splicing sites, right. And what does quantitative transcriptomics tell us? Quantitative transcriptomics tells us about the number, how much transcript is actually expressed, alright. So, mRNA expression using real time PCR actually falls under quantitative transcriptomics, right. Proteomics, I do not need to tell you what proteomics is.

We have covered the entire basic and advanced high throughput proteomics in two entire modules, right. So, this is again just to recap in one line, it will be this branch of study is used to identify, characterize, separate and functionally annotate multiple proteins that are in our system. These actually deals with high throughput analysis of thousands of protein in our body, foods and cells because the final answer is always protein. It is how the protein expresses that leads to normal or disease and healthy function in our body, right. Metabolomics study of all metabolic parameter, we already discussed how molecular diagnostics are used to determine metabolic disorders, various metabolic disorders.

So, the study of amino acid, fatty acid, carbohydrates, what are the important metabolites in our body, those are falling under the purview of metabolomics, right. Microbiomics, very important, microbiome, the intermicrobial flora, so be it on the skin, be it in our oral mucosal system, be it on our gut, every individual has got their own unique collection or signature or fingerprint of microbial flora which can, which does actually determine multiple disease process. So, this is a rising branch of omics also known as microbiome mix where the entire microbiome of an individual is being studied and its relation to various disease pathology are being studied, so it falls under the purview of microbiome mix. So, our job is to bring everything together and to understand everything all at one relate one another, not studying each of them in an isolated manner. So, you already know considering a disease, there are multiple factors for example treatment, right, exposure to any drugs, it can affect a disease process or a pathological process or even a physiological process in many ways, this is a genome, you already know the central dogma, it gets transcribed to mRNA, right.

This mRNA with the study, this becomes the transcriptome, entire mRNA that is present in a cell, that mRNA translates to protein, again that protein can regulate, this is a feedback mechanism that protein can also regulate the expression of the mRNA, right. So, we are studying proteome and that proteome, that protein is actually affecting many expression of other metabolites via multiple enzymes, right and the metabolites can again influence the protein by acting as a ligand and modifying their actions. So, everything is actually interrelated and therefore we are trying to achieve this, we are trying to gather all the information and integrate them and to know in a coherent manner what is actually there. So, before going into the integration process, we have already discussed about big data in our previous lecture, right. So, today we will be using that concept, we will be taking that concept ahead and also be discussing what is principal component analysis or PCA, right.

This will also help us to understand how it helps, I mean it will also help us to understand the integration process. Now, the thing is big data is increasingly becoming the norm, why because more and more R and D is happening every day, big data so much of data, so much high speed you already know how big data is being generated, right. And lots of domains have got their own big data, it is not that if we combine everything it becomes a big data, each and every domain has got their big data and every big data involves multiple variables, right. So, data scientists are continuously working on all of them and they are also trying to analyse those data by developing an individual algorithms, right. So, for example, if I have got big data on my next generation sequencing, I definitely have got my next generation sequencing software to analyse all of them, right DNA sequence.

If I am working high throughput proteomics lab, thousands and thousands of mass spectrometry is happening every day and I have got my software which can analyse all the ion signatures and give us an idea about all of the samples from all of the patients. Now, imagine we have got the same samples from one thousand individuals who are having, who are undergoing next generation sequencing and same one thousand individuals, patients from one thousand, I mean samples from one thousand patients that are also undergoing mass spectrometry. Both are undergoing very high throughput process and both data are big data in such small span of time and they are being analysed by their separate, I mean algorithms. Now if we are to connect them, if we need to draw one is to one conclusion, right and if we need to somehow relate them, it will make much sense to get hold of only the important bits of data that will the most essential and that will help us to understand the each and every sample rather than taking into account every parameter that is being given as an output in all of the individual algorithms, right. Each and every algorithm can give rise to multiple variables, but we should and rather it is makes much sense to only consider those variables which will give us the must know area that can be used to draw conclusions when studying those individual outcome or of the big data.

Let me give you an example, I will give you an example later. So, this is basically the concept of PCA where we only consider those variables that are the most essential that influence the most and we can discard the others while considering the importance of those data. So, what does PCA do? Principal component analysis, this is nothing new, this is actually already incorporated in the software because a lot of information might be there in the raw data, but the software only tells you what exactly you need to know. So, principal component analysis and again a mathematical equation or statistical thing which extracts in this case PCA is generalized term, it can be used in many things, it extract the most important information, right. What does it, why is it essential? Because it leads to compression.

So, huge amount of data a terabyte of data, I just give you an example, right. You are only collecting valuable information. So, it may be so that 1 terabyte of data is represented in just 1 gigabyte of information that is a huge reduction, right. So, less important information are discarded, we have only few data points to consider and total analysis becomes very easy, description of the whole data becomes very simple and becomes very easy to analyze the data set. So, we always try to extract only the most important information, this is very helpful, very cost effective and very time saving.

So, PCA however, you must be having a question then what about all the other information, right. The could there be any process that I am only considering few data points some data which may have been important is lost, answer is yes. If you are having this query you are absolutely right. So, we are losing some data at the cost of analyzing more and more data, right. If we try to analyze each and every data point it will take years and years of time which is not possible in high throughput analysis.

So, basically it is a trade off, right, between faster computation and less memory consumption versus information loss and this process has to be standardized, it has to be min maxed to reach at a point where with all the minimal data we have called the vital information. So, basically it is considered one of the most important tool of data analysis and PCA this principle component analysis is one of the key components of analyzing big data which helps us to integrate all the multiomics data because each and every multiomics data set is an enormous data set, right. So, we need to filter out the most important information and then integrate them together. So, this is a very simple illustration of PCA.

See here are multiple fishes. So, each and every fish have got definite height, have got definite weight, suppose there are tens of thousands of fishes and each and every fish have got their important data point. So, it is variable, right, height is a variable, width is a variable. Now we have got information about tens of thousands of crores of fish and it will give rise to a very big data set. Now we have to analyze each and every data if we just try to search single fish to have an idea about the pattern, right. However, you can see, so one group of mathematician what they have done, they have composed, they have plotted the height of the fish and the corresponding width of the fish.

So, basically they have made a correlation, right. So, and that gives us a shape score. Now this correlation, this slope, this shape score we can see since most of the fishes that are more in height also more in width and the fishes that are less in height are also less in width. So, in this case just by I am giving a very simple example for easy understanding, we have got another variable, basically this is a slope which can now determine considering all the mean information is the most important variable. So, if the we consider this, if we have this information then just by calculating one parameter, just by calculating one variable we have got information about each and every type of fish, right.

However, there will be, there must be some exceptions in which the fishes that are actually very long that are not much in width and there might be some fishes that are actually very wide, but they are not much long, they are also there, right. However, they become less important parameter and that when we are I mean summarizing this data in order to integrate them or in order to draw conclusion, we will ignore this other data points, but we will consider most of the fishes that are in the tank because they fall under this mean slopes core, right. Now this might seem very abstract and far far away from this is standard, right. So, this might seem very far away from what we are trying to discuss, but believe me just noting this one single variable has saved multiple multiple millions of terabytes of data. Now you can consider I will give you one example of real world application of PCA.

You I last day I told you where we are discussing big data and AI and ML that is facial recognition, right. You know you just put your face in front of your phone and then it happens the phone unlocks, right. That is one example where the machine is learning your facial features. However, that same function is often there in biometric system. So, that is known as face unlock biometric whenever you are entering into a building you need to just present your face in front of the camera and the camera will recognize that you are present and that will mark your in entry.

You may have to do the same when you are out doing that is also outgoing biometry. Now this facial recognition software is a very small hard drive that is present in the machine and it has registered more than tens of thousands information of tens of thousands of employee who are actually getting in and out and it promptly recognizes all of them. Now imagine if it was a very high definition camera that maps faces 1 is to 1 and each image is around say 5 to 6 MB megabyte then you can imagine tens of thousands of image will take so much volume of data, right and there will be the device will be easily filled up and out of memory, but it does not happen why? Because computer algorithms have refined that with very minimal features we can actually use recalculate back the information about the exact facial structure. So, even if we are losing important information, right exact details whether I have any freckle or wrinkle over here the basic information that this facial structure is mine the machine is able to tell us. So, this is one of the very important application of PCA which is extremely helpful in reducing the data load in integration of multi-omics data.

So, one very important concept is principal component analysis which is actually incorporated in every software. Now mind it as and when we are studying every omics and we are talking about integrating the data mind it only the principal components of the data are actually integrated not everything, fine. So, basically is an overview of multi-omics. So, what we are trying to do here we are trying to map a phenotype with genotype, right. For example, in case of whenever we are studying the genotype this is basically the picture where all every type of omics have been represented.

So, when we are studying genomics it can be single nucleotide polymorphism, cnv copy number variation, LOH loss of heterozygosity so on and so forth genomic rearrangement, rare variants everything falls under genomic study. These are the various studies that we are looking for under epigenome, right. So, this is actually considering a single sample, right. We are we have done results of a single individual or a single sample or single cell or a single animal sample. And we are studying the various processes at every step at epigenome level we are looking for DNA methylation, histone modification, chromatin accessibility so on and so forth, right.

In transcriptomics we are looking for any gene expression alternative splicing, in proteomics we are looking for protein expression, we are looking for the cytokine array and finally, at the level of metabolomics we are looking for the metabolites that are present in plasma or not, right. So, we are trying to fit each and every process one is to one. Suppose we are targeting the same gene, we are targeting the same DNA, then gene, then mRNA, then the proteome, the protein and then finally, the metabolic product. So, we can actually do the analysis and fit them horizontally or vertically I mean interface them to understand what is actually going on. Now how do we do it? The methods for data integration basically can be two types.

Number one where each is placed is done one after another, right. So, different data types in a stepwise, linear and hierarchical manner. So, first we are doing the genomic study with that we are now doing the transcriptomic study, epigenomic, then transcriptomics one after another and we map all of them, right. Or we do everything and consider all of them together. Both approaches are there and not every approach is foolproof and it depends from model to model, disease to disease, disease case to case which model will be best suited for integrating various multinomics data.

Now the first stage, first we will be discussing multistage analysis, then we will be going to multidimensional. Number one, equatorial expression, quantitative trait, loci analysis. So, what is this? Mind it we are always trying to map a genotype with phenotype whether something is the there in the gene and that has expressed in protein level or not, right. So, what do we do? We first associate multiple single nucleotide polymorphism with the phenotype and then filter by the significant threshold whether this phenotype is. So, depending on our need or the research question or depending on what do we want, we can actually set what type of phenotype we will be accepting and what type of phenotype is qualifying eligible for the change, right.

So, that is called the significant threshold. So, basically depending on that amount of change we can consider this phenotype will be eligible for the genotypic study or not or whether this genotype will ultimately result in the phenotypic expression or not. So, first the SNPs will be associated, alright. Next the SNPs that are associated with the phenotype they will be tested with other omics data, fine. For example, we will check for the association with gene expression data and also methylation, metabolite as well as protein. So, in every case we will be looking for the expression of that disease process, right.

Next what do we will do? Next we will be testing the omics data that has that we got in the step 2 and we will be correlating. So, what information we got in the earlier gene. So, whether it is actually correlating with the phenotypic expression or not, right. So, we will be considering all the genomic data and then we will be considering with the proteomic data and we will draw a correlation line. Now it is very important there are two types of eQTL, number one is where the expression of the protein is affected by a gene that is remotely located that is known as trans.

And where the effect of the phenotype I mean the phenotypic expression is controlled is regulated by a change in genomic phenomena or the gene nearby gene that is known as cis expression quantitative trait loci analysis. So, basically we are trying to map them one is to one this basically very if you find this language very difficult just know the concept. We are trying to map the phenotype with genotype, we are studying various we are doing various genomic experiment, we are basically we are studying the single nucleotide polymorphism that are associated with this phenotypic expression. Then we are doing the genomics testing, we are doing the proteomics testing and then we are trying to relate them one by one by drawing various correlation curves and then we are trying to match whether there is any correlation or not.

This is basically the thing in simplified language. So, this is one of the very important and this is all done by software and mind it this is not a single we I told you one single case this is being done for multiple lakhs and lakhs of genes that are being studied together and lakhs and lakhs of simultaneous phenotypic expression that have been reported with those genes. One step ahead right in multi stage analysis what do we do? We consider allele specific expression. Now in diploid organism right in haploid organism we just analyze one genomic study whether it is nearby or how far away. So, depending on the cis or trans e QTL for diploid organism we also need to consider the fact there are two alleles that can express. So, one can remain dormant, one can remain active, one can be functional, one can be non functional.

So, basically it is similar to analysis of expression quantitative loci single nucleotide polymorphism, but here we have to also correlate the similar I mean alleles alright. So, what allele has expressed that additional data is also considered right. So, what does it do this test whether maternal paternal allele is preferentially expressed or not right. And this one associates this allele with cis element variation and epigenetic modification. So, what do we mean? Suppose we are trying to find an answer to a question whether this genetic expression has led to this phenotypic expression or not alright.

And now there are two possibilities that there the genotypic expression or the genomic variation or genomics S and P are possible in one allele or the other right. So, we need to first find consider that possibility, then we need we can associate whether the element that is happening to the nearby gene. Generally allelic expressions of such integration of multiomics data are done by cis fall under the cis-EQTL model of integration of multiomics data alright. In the next step that is domain knowledge overlap what do we do? So, mind it one step I mean it is one step going over the top of another right. So, all that we are doing in EQTL are also being considered in allele specific expression right.

What this do? An initial association analysis performed so by the S and P gene expression variable right. Now after the whole analysis is done alright there is an additional step that is known as functional annotation where with the help of software suppose we have demarcated the entire biological process that is happening in a control group and therapeutic group. Then the various processes are actually functionally annotated suppose we have got a genomic analysis of a pathway we have got transcriptomic analysis. So, at every step every step involved in a pathway every gene involved in a pathway every DNA involved in a pathway are analyzed separately. Suppose expression of I mean insulin right so insulin is generated it binds to receptor there is tyrosine kinase phosphorylation and finally, it results exerts a function.

So, every time each and every DNA that is responsible for each and every protein is first analyzed then each and every mRNA is analyzed then each and every protein is analyzed in every step the function is annotated for that we know right. Then we have got a picture where we have got an entire genomic sequence with functional annotations, we have got a transcriptomic sequence with functional annotation, we have got a proteomic sequence with functional annotation right. So, then based on their functions all the data are tried to merged. So, overlapping of the function the knowledge we already have based on that so what area is acting as a receptor. So, this DNA is for the receptor this protein is for the receptor and this MRN is actually for the receptor this total knowledge or total data of concerned omics elements that have got similar function are actually overlapped.

So, what does it do this approach enables the selection of association result will functional data to corroborate the association. So, basically it might be so that due to some change the protein has differentially expressed right there might be some change in the receptor DNA or there might be no change in receptor DNA, but there might be change in receptor mRNA or there might be not any change in receptor mRNA, but the receptor protein, but we know the problem is in the receptor. So, once we have got knowledge that this is the functional domain that does this function if we integrate or if we look at the data by trying to overlap all the functional knowledge then it is much easier to achieve at a conclusion that this is the actual variant in the disease process. And then there is another approach where everything now here we were considering everything one after another. So, be it allele specific, be it expression quantitative trait loci or be it multistage analysis we were taking each and everything one step at a time and then we are trying to merge them.

Whereas, what does multidimensional analysis do multidimensional analysis as I told you we already have got the data huge data from multiple domains. Now, our job is to combine all of them and then achieve at a conclusion. So, everything is dealt simultaneously everything is read analyzed simultaneously. So, this multidimensional analysis can be divided into three category.

Number one, concatenation based integration. Now, for those who have some knowledge about computer programming concatenation, string concatenation means simply joining or combining. So, if you type hello and world and if you type hello plus world you give a print function that it will read hello world right. So, this is the basic very basic function. So, we are trying to combine everything. So, multiple raw processed data is actually combined into one single information and then it is analyzed right.

So, it is one model right it may not be perfect, but it is very important that the total that by one single analysis since every data is combined. So, data from S and P matrix. So, we have got thousands of patients and each of array and whose S and P study was done mutation or anything was done. Then we have looked into multiple gene expression, we have looked into the mRNA matrix and now we are trying to combine all of them the phenotype and genotype. So, all of the data are first merged and then we look I mean we design one variable by which we need to study that.

So, what is the problem here? Basically the number of measurement is less compared to the number of samples. So, this is there is one issue that is known as inflation of high dimensionality. So, what happens? Basically the data is huge right and we are measuring the data in one measurement. So, this is basically a challenge that we need to design one analysis method by which which will be able to take care of all the information all the necessary information that is present in this data. Otherwise what could have been done? We can filter out various important information from each and every data, then we can analyze mind it in every step PCA is already done.

Do not think that ultimate petabytes of data being analyze integrated together and then we are studying. The most important information from this S and P matrix is taken into consideration the most important information from this other gene expression matrix is taken into consideration so on and so forth right. So, only this model deals with combination or concatenation of all the important data and then analysis of them. What is transformation based integration? Here also we have got three different types of omics data right.

So, we are trying to map each and every data sets before analysis. However, since this model considers that there are three different types of data that are there. However, we are only considering only one variant of data set to analyze. So, here three initial graphs are all spanning trees, how spanning trees are each and every node and how each nodes are associated. So, they generate type of this type of data these are known as spanning trees. So, each one of them is representative of a disease function right of the same disease.

Now see one of them the best suited model is actually selected and it is analyzed when we are trying to map the phenotypics to the genotype right. This modeling approach is then applied the level of transform matrices. So, basically here what we are doing we are combining all of the data and then we are analyzing. Here we are trying to transform or physically map all the individual spanning trees into one type of model so that we can analyze only one type of data. So, basically this one set of data has got information from all, but we have designed it so that only one model or one variant of spanning tree has been selected.

In contrast to this, this one that is the model based integration this considers analysis of all analysis of all the individual data and then drawing information it draws a resultant model from all of them right. So, mind it you please do not think these as individual antibodies or structures like that these are all data sets we are trying to integrate or analyze data right. So, it is how data is an this is an inform representation of type of data only right. So, here we have got three different data sets each and every omics data have generated three different type of data which has got which is huge data in themselves and then the computer algorithm has designed a resultant data set a pattern which has considered all of the individual data sets and again the number of analysis is of a resultant data set. So, mind it everything that is being done the arrows are basically computer algorithm that are filtering out excess data at every step as a part of the principal component analysis.

Now, similarity network fusion SNF what is this? Now, this is basically how we map all of these. So, how do we get these right? So, these are I mean you can see suppose I mean figure one this is a representation of mRNA expression DNA methylation sets of same cohort of patient we have got say 30 or 50 or 100 patients whose sample we are analyzing right. So, this one is a data of mRNA expression and this one is a matrix of DNA methylation the color coding is dependent on our result that is the what we have set. Now what is the aim of all of this? So, similarity network fusion this tries to aim discover subgroup clusters why do we need to discover subgroup clusters mind it from previous class we are trying to design an unsupervised learning of ML model finally, all of this data will be fueled into machine learning training right and how we are basically studying how do we get those data so, that they can be trained by the machine.

So, basically we are trying to achieve this subgroup clusters of non level data. Now see mRNA expression and DLM methylation sets of the same cohort of patient now patient by patient similarity matrices for each type. So, we have considered here just by PCA PCA for component analysis each and every patient whose DNA methylation has got some resemblance or some information that is similar to that of the mRNA expression changes. Suppose this patient A has got one amount of mRNA expression and has got a same or has got a different amount of DNA methylation. Now patient B can have same amount of mRNA expression and different amount of DNA methylation compared to patient A.

Patient C can have different expression, but same DNA methylation compared to patient A. So, basically this is a similarity index of multiple patients that we have done. So, mind it this is just like the diagram of the fish with height and width here multiple patients are being done. So, each and every patient have got their own DNA methylation study and DNA mRNA expression study. Here we have prepared a similarity index alright. So, we have pulled we have grouped the similar data in same matrix right just to have just to understand the person with this amount of DNA methylation had that correlating amount of mRNA expression.

So, patient by patient so, for each and every when this is achieved then the software can generate patient by patient similarity network which is equivalent to the patient by patient data. Now patients are represented by nodes. So, these are individual patients right and patient similarities are represented by edges or the connecting lines. So, this is how a big scattered data can be pictorially represented in hugely concentrates or makes the representation coherent for the data scientist to analyze.

Now this does not end here. So, now this network fusion this happens iteratively it keeps on up it is a repeating thing that updates each and every network to similarity of information based on the amount of disease compared to other networks and thus making them more similar with each step. So, more and more patients more and more similarity networks will be found right and this iterative fusion leads to convergence in the final fuse network. So, basically the depending on the type of similarity for example, we are denoting the mRNA similarity with blue line. So, this is the diagram of mRNA similarity and we are denoting the similarity of DNA methylation with pink line. Now when we are combining them we will definitely find a pattern and this is what is going on over here.

So, for example, you can see there is no similarity in DNA methylation network. However, in here there are similarity in sorry since the blue is mRNA there is no similarity in mRNA based expression, but there is similarity in DNA methylation. However, when we are repeating more and more like thousands of data definitely at some point of time there will be similarity or we will map those in such a way just by turning the whole thing this is a figuratively figurative expression we are finding similarity. And this is basically representation of how similarity in data is there it is hugely I mean it is a huge achievement. So, that we can map everything together by such minimal figurative representations. Now this is one real world example of the how it looks in software for example, here 215 patients with glioblastoma is a variety of cancer neoplasm that is represented by similarity matrices this is how it is happening right you can see the similarity matrix over here and then SNF alright.

Similarity network fusion has been done and then it is much easier for the machine to find subgroup cluster. So, mind it the purpose of all of them we have done SNPs, we have done DNA methylation, we have done mRNA expression of thousands of patients. Now we are trying to find patterns in the machine learning one very important thing is finding patterns and associations and this is how it is achieved right. So, the clustering representation reserved for all four networks to facilitate visual comparison. So, definitely there are everything has been so, micro RNA mRNA and DNA methylation has been studied over here.

So, similarity networks all have been plotted and then it is much easier to find clusters and subgroups for machine learning training or scoring right. So, finally, this is the combined similarity index that has been achieved by computational software. So, next we move on to a very important topic that is factor analysis it is similar to principal component analysis. So, what happens? data from multiple matrices are achieved and they are overlapped. So, basically again we are trying to integrate everything, but we are only including the most important information from each and every omics data very important.

So, suppose we are trying to find out a factor that is related to inflammation any factor right. Now this higher inflammation can be related with the higher expression of certain genes for example, interleukin gene of IL-6 or it can be associated with expression level of higher proteins or it can be due to inflammation of increased amount of metabolites. So, we have got information from all of them these matrices are individual studies of multiple cases. What do we do next? We combine all of them. So, this multinomics factor analysis it this term is known as decomposition basically merging of all the three types of matrices in a common matrix by finding similarity all right.

So, you see white cells in the markers correspond to 0 that is inactive features. Now everything this again this matrix has got information about all of the patients right, but the final data is so less compared to all the other individual volume of individual data. So, we have done assays for proteomic study, genetic study as well as transcriptomic study and or proteomic study, metabolomic study and genomic study, but the final results are specifically telling us which domain we need to focus more. So, by combining all these factors this multinomics factor analysis helps researchers to understand what are the biological molecules, how they are interconnected basically what are the main drive. The main drive is why is the variation across all these samples. Now by combining all of these we do have a convincing answer that in this case there is genetic variation that has led to the metabolic variation that has led to the proteomic variation.

So, we are always trying to find what is abnormal or what is the variant in a multiple disease process. So, a similar disease where multiple patients have present to the different types of inflammatory response this helps us to understand the multinomics data. So, basically the thing is it can be visualized as a principal component analysis. So, in principal component analysis correspond to any data we are trying to gather the most important information. Similarly, this model of multinomics factor analysis considers certain variables which are important for you to know if you are trying to design number one it will it has to consider the variance decomposition as saying the multiple proportion of each variation that is there in each data modality or each assay how it is varying then it will I mean compare all the variance and it will merge then semi automated factor annotation based on the inspection. So, basically semi automated means there is some manual component where we need to know and we need to annotate what are the factors that might be present.

So, after semi labeling the data the machine will then annotate the other levels of data and it will be combined together. So, there are multiple components to each and every matrix. So, each and every matrix needed to be treated differently and there are data scientists that are first labeling or annotating each and every matrix and then by achieving at a visual result they can be easily combined. Now one thing there might be so that there might be some failure in any matrix suppose this assay has failed right.

Now there lies the importance of machine learning one thing is imputation or refilling of lost data. So, imputation of missing values including missing assays it is much easier to do if we have got information about the other matrices. Suppose we in a single patient we do not have information about what changes happened in their genetic assays in their one single experiment might fail, but if we have got information about their mRNA expression and finally, protein expression and by combining all of them weak and by knowing the pattern how it happens we can easily impute the missing results in their assays. So, that is the in short and in brief the very you can say in a lucid language the thing has been discussed, but the thing is actually very complicated when we are trying to analyze each and every data needs immense training in data science to achieve all these facts. However, what can we say about the final outlook and viewpoint? The thing is the results of all of these omics data mostly comparative. So, we are taking a matrix and we are doing a disease for suppose insulin resistance and for insulin resistance we are doing this study.

So, all of the patients mind it though ideas to find variation all of the patient have got similar disease process, but each and every one are giving various results and now we are trying to find answer where lies the problem right. So, we do have some omics data from healthy and diseased individual and assume that the difference is directly related to the disease right for all the models all of the softwares, but one very important issue is health and disease never go handy then they are two very different heterogeneous groups and there are many confounding factors such as population structure, the cell type, various composition bias, batch affections there are so many unknown factors that actually might lead to problem right. And this is one of the challenge which we should be cautious about. So, how we are actually trying to analyze or what we are trying to achieve with this omics data is very important. For example, sex is one of the major determinants of biological function and there are multiple diseases or even multiple physiological features that show sexual dimorphism and therefore, any personalized treatment approaches mainly the cornerstone we are trying to achieve this multi omics data integration whatever information we are doing or achieving we are trying to get into that precision medicine domain where we know we have so much information about any disease process that we are able to devise any drug or any therapeutic approach that will be effective.

However, sexual dimorphism drastically alters one important I mean the observation that has to be taken into account. So, differentiating causality from correlation based omics analysis already mentioned open question. So, we always try to see this has happened and this is the result, but why this has happened the result might be lying in omics data, but we need to find it out. So, but what are the benefits of multi omics data? Very important compensate for missing and unreliable information on a single data type. Suppose one variation of genetic gene study has not been done, but we have got information on the transcriptomic as well as proteomic data.

So, one protein may be easily studied. So, now, with the help of this multi omics data integration technology we can have we can retrospect we can predict even what will be the outcome of the genetic testing even though it has not been done in one case. So, multiple source of evidence points to same gene or the same pathway to one can expect the likelihood of false positive is reduced. So, we diagnose the disease or we see that the patient is positive or has got a features at genetic level, transcriptomic level, proteomic level, metabolomic level then. So, if we are trying to design a multi omics test. So, that since it has got multiple data points of analysis that result result of false positive means the test being positive in the disease person not having the disease of the condition is very low.

And very important it is likely that you can uncover the complete biological model considering different level of genetic and genomic and proteomic regulation are there, but not everything is answered right. So, unraveling of one will lead to easy unraveling of the other. So, lastly after knowing all this we should know what are the challenges. Number one challenge you will be able to predict this if you have not predicted by now number one is integrating this complex genomic data with traditional clinical work flow.

So, complex traditional clinical work flow patient outpatient sheet or data sheet prescription truck history everything. So, we need to integrate everything together. So, this genomic data along with the qualitative data or the patient data of treatment has to be interface somehow. So, that they can be integrated then one information might lead to another. And not only integration the complexity of multidimensional data analysis as well as the data interpretation these two are the biggest hurdles till date of multinomics technology, but as I told you the challenges are the ones one area which always love out because it is the most researched area and in future maybe in 5 years one challenge we might become one benefit. So, basically we are trying to address the regulatory hurdles and data privacy concern that is always there I told you what are the ethical concerns in molecular diagnostics.

Very important education of healthcare professionals by changing their approach from traditional medicine to precision medicine approach is a big hurdle. And there is resistance in the community and that has to be taken up by the changing the healthcare system and the provider community they need to be responsible also the cost. It is not easy to design multiomic experiment or multiomic high throughput testing for all diseases right. Therefore, managing the cost of all these technologies and balancing all of them so that final affordability is reached.

So, that can be sustainable is one of the challenges in omics technology. Also ensuring equitable access to advanced diagnostics and therapies and navigating disparities in healthcare as I told in the last class everyone should have access not only the privileged or the well to do so all classes of socioeconomic patient should have access to all of this. So, that it can be globally implemented and precision medicine practice can be adopted by all centers. So, to summarize we have learnt about and we have learnt in brief I would say we have touched the surface of omics data and the scope of omics data we have recapitulated what are the various types of omics approaches and how they are interrelated. We have discussed in very superficially what are the various models to integrate multiomics data and we have also discussed what are the benefits as well as challenges and future directions of omics technology. So, these are my references for today's class and I thank you all for your patient hearing.