

Interactomics: Basics and Applications
Prof. Sanjeeva Srivastava
Deeptarup Biswas
Arijit Mukherjee
Department of Biosciences and Bioengineering
Indian Institute of Technology Bombay

Lecture – 56
Secondary Data Analysis

Hello students. We have talked about many technologies both label based and label free platforms. And what you may appreciate that omic tools often generate huge data as you have seen in the case of microarrays, SPR and NGS platforms. So, data could be obtained from different equipments; sometimes when we are doing measurement using microarrays, then you are measuring the intensities. When you are looking at a mass spectrometry type of platforms then you are looking at area under curve.

So, different instrument gives different ways of measurement. But finally, what we get we get huge data set and then what we have to try to do, but further from that how best we can make sense of this biological data set, how much it is relevant to address the primary question which we are interested to address. In the previous lectures you were exposed to the basics of a statistical analysis and the importance of having clear understanding of the a statistics, that are to be implemented to get the meaningful insight.

And interpretation of your data the data obtained from these high throughput omic technology platforms, mostly comprised of huge range of values that need to be scaled down to perform the analysis. Also the big data that comes with lot of you know variability, sometime it is day to day experimental variability, sometime instrumentation error batch effects of course, manual errors will be there as well. So, therefore, there is a need to look at how best to normalize there data right. And how to remove or minimize the technical artifacts to obtain the biologically relevant and meaningful data set.

So, we need many pre processing steps, which includes data as scaling normalization before we actually proceed for the a statistical test. In today's lecture at hands on session I have two of my research scholars Deeptarup Biswas and Arijit Mukherjee. Who will walk you through

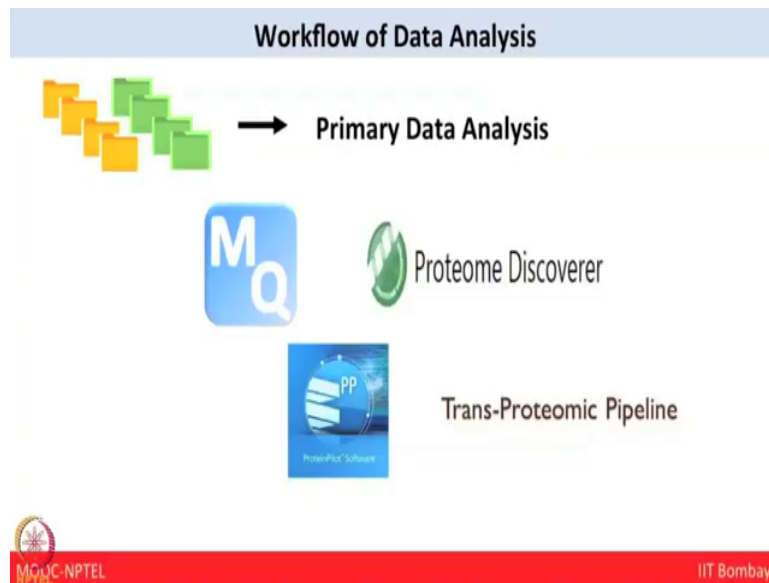
different steps which are required for the analysing big data sets; they will further elaborate on different aspects of a statistical analysis and data visualization and plotting. I must would like to mention that there are many very nice open access platforms and software's available now.

And good idea for all of your students are participant to explore many of these software tools to make best sense of your data. Not only how to best analyse the data, but also how to interpret data. And finally, how to effectively present the data and this is where even data visualization becomes very crucial. The data generated from these techniques comes in the form of big tables, but big tables does not convey the biological sense of what we want to address. Therefore, you have to convert all these findings into the graphs and the visual way of disciplines.

So, that therefore, we thought you should be learning some of these software's and tools which can help you to best utilize and visualize your data sets. So, in today's session we will have the basics of secondary data analysis and it will walk you through different steps involved in pre processing of the data, basic statistical analysis data, visualization and plotting. So, let us start today's lecture.

So, let us take an example, that we are we have designed an experiment, where there are 13 controls and 15 disease samples. So, after running all these samples individually, we got 28 files of which 13's are control file and 15 are diseased files. So, now, we need to analyse these files and need to know, how the controls are different from disease. So, the first thing we should do is, we need to take this files and do a primary analysis.

(Refer Slide Time: 04:33)



(Refer Slide Time: 04:33)

Workflow of Data Analysis



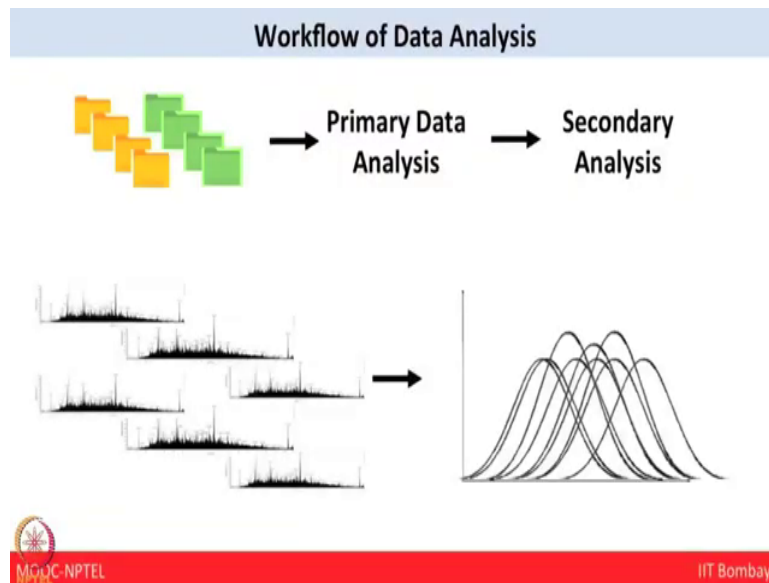
The screenshot displays a software interface with a table of protein data analysis results. The table has columns for 'Protein', 'Protein Group', 'Peptide Group', 'PSMs', 'MS/MS Spectrum Info', 'Input File', 'Specialized Traces', and 'Result Statistics'. The 'Result Statistics' column includes 'Thres. Min.', 'Thres. Max.', 'Thres. Avg.', 'Thres. Std.', 'Thres. Min.', 'Thres. Max.', 'Thres. Avg.', and 'Thres. Std.'. The table lists 28 proteins, each with a unique identifier and associated data points. The interface also shows a status bar at the bottom with the text '5210432 Proteins, 521 Protein Groups, 1677 Peptide Groups, 2183 PSMs, 8533 MS/MS Spectrum Info, 12 Input Files, 2 Specialized Traces, 174 Pe...'.

#	Protein	Protein Group	Peptide Group	PSMs	MS/MS Spectrum Info	Input File	Specialized Traces	Result Statistics
1	AAITSDLEALGR			1	2	2	Q1593	0 1409.71
2	QGANHER			1	1	2	Q1585	0 1014.53
3	QGANVITSDVANDPALK			1	8	2	P0603	0 1011.54
4	QVTPADVDSGNPK			1	4	1	P13787	0 1408.75
5	QFAPEYK			1	2	2	P13627	0 1011.47
6	QEVDSQPSIAR			1	8	4	P0709	0 1516.70
7	QEPSGUTTTAVYSLR			1	3	2	A0A28VEY1	0 1029.54
8	QEPERNECTLQNDKDFLPR	1-Celastrolinethyl ECE	X	1	4	3	P0709	2 2036.22
9	QEPERNECTLQHK	1-Celastrolinethyl ECE	X	1	4	1	P0709	1 1714.79
10	QEVSLFNAFGR			1	1	1	Q26HE7	0 1300.72
11	QELAALEK			1	3	1	C3A25	0 1014.55
12	QSPQNAVNTDQK			1	2	2	P0275	0 1308.61
13	QGDGKTLR			1	6	4	A0A7GCPW6	0 1103.53
14	QDGFVGR			1	14	2	P0709	0 1111.88
15	QDQDQDFGR			1	2	2	P0275	0 900.44
16	QAFIENIESEYDQYK			1	3	4	P0403	0 2141.01
17	QAVTPNPTFYAK			1	4	2	P0846	0 1062.71
18	QAVGAGLPIRDTCTTK	1-Celastrolinethyl ECE		1	3	1	P0473	0 1541.54
19	QASDPLK			1	2	2	P0406	0 829.44
20	QAGEYALLAK			1	2	2	P0570	0 1419.74
21	QATZAVISR			1	1	2	Q0H84	0 1100.64
22	QASPSVPR			1	1	2	P1011	0 988.56
23	QASLQGARLK			1	1	2	P0846	0 1221.65
24	PIPHVAVR			1	1	1	PIPHVAVR	1 1000.17

So, after primary data analysis what you will get total number of files which we took an example of 28. When we are talking about all 28 files, there is number of things that come into play. Like technical variation, experimental variation, batch effect and all this kind of thing; that means, we need to apply some strategy or to normalize the data. So, next part of the data analysis is the secondary analysis, where we first transform the data that can be locked to transformation.

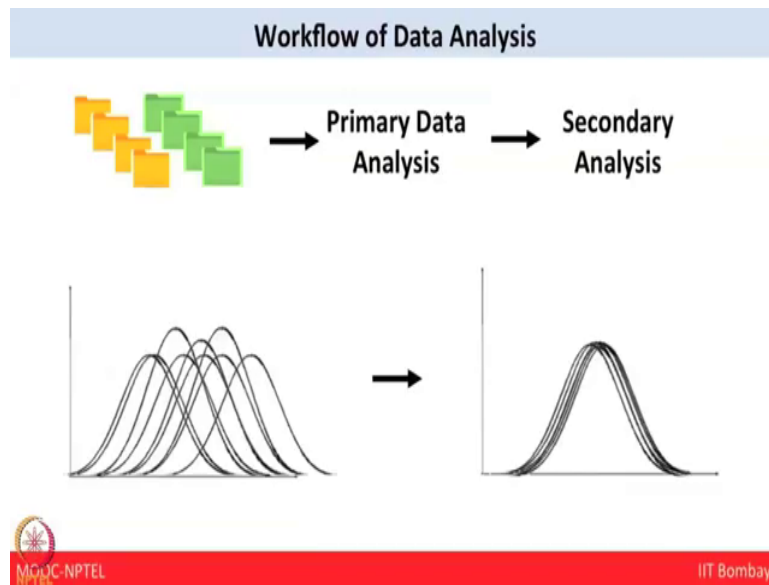
And try to understand whether the data is following a Gaussian distribution or normal distribution or not. Followed by if dataset is skewed, then we go for sample wise normalization and if needed we do data filtering and followed by different statistical test.

(Refer Slide Time: 05:28)

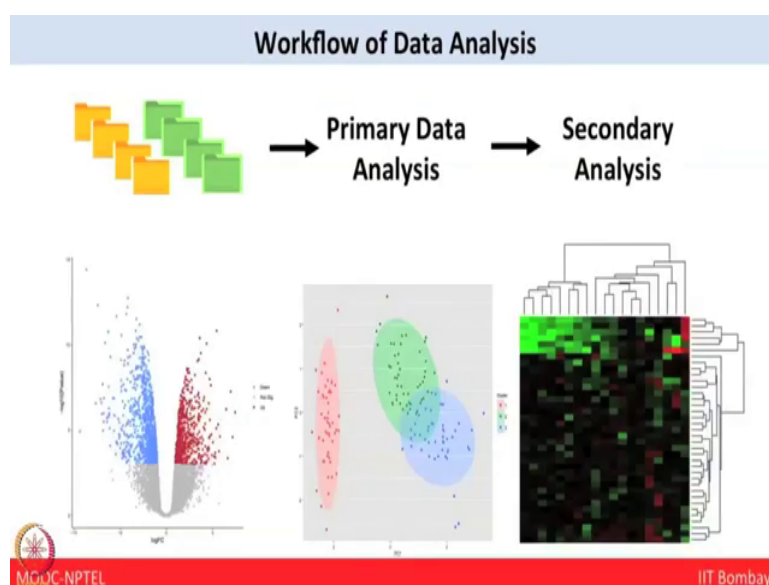


So, after the analysis of the 28 files, what we get? This kind of distribution and with the help of the normalization transformation and scaling, we need to get the data like this.

(Refer Slide Time: 05:35)



(Refer Slide Time: 05:40)

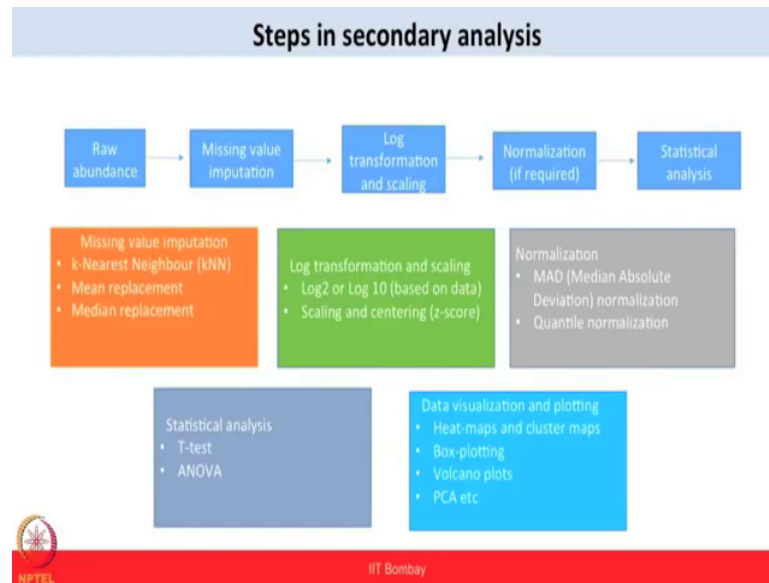


So, that we can proceed for define statistical test and statistical plot. And finally, after the secondary analyses we try to link our significant protein list with biological pathways protein protein interaction and different pathway enrichment networks to understand the biology behind the disease. So, now, let us welcome Arijit Mukherjee a junior research fellow in proteomics lab, who will be giving a detail information of secondary analysis.

You are not aware of the overall overview of the complete data analysis workflow. So, in data analysis, we go through 3 steps mainly; primary analysis, secondary analysis and then to the tertiary analysis. So, what you get after the primary analysis is the list of masses with their respective abundances or intensity. Now from these respective abundances of intensities of a list of proteins so, what we have in thousands of proteins we need to make valid inferences

based on some statistical tests. And this statistical test and other forms of transformation scaling, normalization this steps comes under the secondary analysis steps.

(Refer Slide Time: 06:59)



So, let me take you through the steps in secondary analysis. What we did after the raw intensity values, we do missing value imputation. So, missing value is a common feature in omics data set; this may occur due to some technical limitations or due to some manual handling. So, we will be talking in detail about this point about missing value imputation in terms of k NN that is k nearest neighbour mean replacement median replacement. Next step is to do log transformation and scaling.

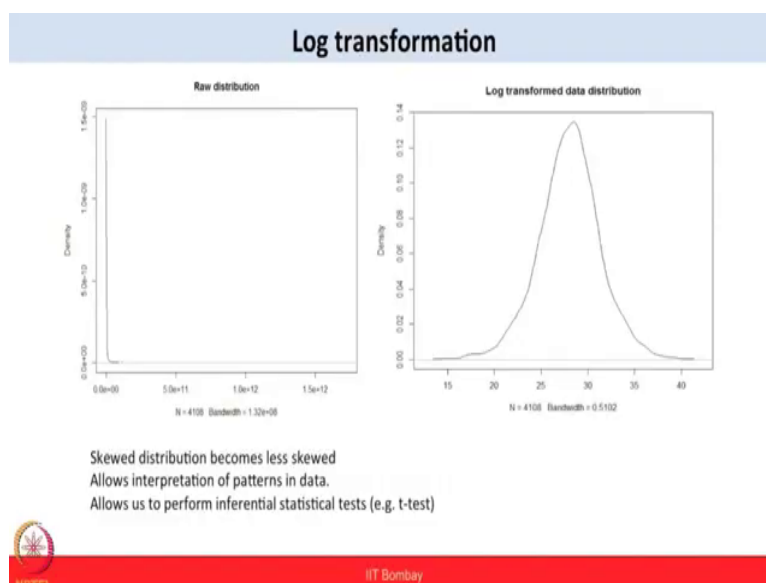
So, we do generally log 2 or log 10 based transformation and scaling and centering, in order to be able to compare different kinds of distribution. Because different samples may take up different kinds of distribution after log transformation and scaling. And the next step we can

perform is the normalization. Normalization comes under median absolute deviation normalization or quantile normalization. There are many other kinds of normalization that depends on your data structure that we can perform.

But here we will focus on these two kinds of normalization specifically. And next comes the statistical analysis steps. Statistical analysis is performed to test the hypothesis basically. So, depending on your experimental question, that you want to know the significantly desirable that are genes or proteins or mRNA's that is present in digit sample has compared to control are controls.

So, we need to perform T test or ANOVA test or regression analysis, these things will be covered in this lecture and finally, comes the data visualization and plotting. In data visualization and plotting, we can draw heat maps or cluster maps, box plotting, volcano plots, etcetera. So, this is the overview of the lecture that is going through to a secondary analysis that we are going to talk about.

(Refer Slide Time: 09:01)



Now, coming to the first step, after you get your raw spectra if you look at the mass spectra abundances; it comes in the terms of e to the power; e to the power 9 e to the power 10 and so on. So, from these numbers it is quite difficult to play around this and to make valid assumptions. And as you can see in the figure on the left side, that with a 4108 elements that have been plotted on the in a frequency distribution curve this is quite skewed. And this kind of skewed distribution we cannot make any statistical inference based on this distribution. So, what we do, we perform some log transformations; generally we take a log base 2 or log base 10.

And transform the data on the right hand side of the figure, you can see the log transform data distribution; it is not perfectly normal distribution, but it is assuming to be a normal distribution. So, this log transformation brings the data range to a workable range that is 15 to 40 in this example. So, these numbers are quite easy to make inferences with this kind of

data. So, coming to the importance of log transformation, that as you can see in this figure that the skewed distribution is transformed to normal distribution or becomes less skewed.

And this allows you to interpret patterns in the data. And always we can perform some inferential statistical test that is T test or ANOVA etcetera. In as you know in t test the basic underlying assumption is that your data distribution should be normal distribution. So, this allows to help us to perform such inferential statistical tests [vocalized-noise.] And in omic studies which generally perform log 2 or log 10 transformation that depends on the objective of your experiment. In biology generally we do log two transformation.

Because we are interested with genes or proteins are in the fold change of twice; as compared to the normal or dysregulated or downregulated at a fold change of 2 values. In that case it depends the scaling whether you take base 2 or base 10 that depends completely on the objective of your experiments. Now next coming to the topic of scaling and centering; if we take the log transform distribution it looks like a normal distribution. But when let us suppose we are taking control and disease samples the control samples will have a different mean, if we log transform that values.

And the disease samples will have a different mean if we log transform the values, but this two distribution; however, it is normal we cannot compare the two means when the two means are very different we need to scale and center. So, that the distribution takes up the similar mean or median values and they are distributed around those mean or median values. And this is the purpose of scaling and centering that we can compare different samples; since all samples become normally distributed with mean of 0.

(Refer Slide Time: 12:12)



Scaling and centering

- Scaling and centering is performed generally in terms of z-score.
- The z-score can be computed as $z = \frac{x - \mu}{\sigma}$ μ = sample mean
 σ = Standard deviation

z-score ranges from -3 to +3; assuming a normal distribution.

Other methods:
Scaling in the range of 0-1.

Score = $\frac{x - \text{min}}{\text{max} - \text{min}}$ Min = Minimum value of all observations
Max = maximum value of all observations



And this is performed in terms of z score the z score as we can see in the formula is the x minus μ upon σ x is the individual values of the variable, μ is the sample mean and σ stands for the standard deviation. So, this z score reflects distance from the mean in terms of standard deviation. Assuming a normal distribution the z score should range in minus 3 to plus 3; because in normal distribution 67 percent of the population lies within the mean plus minus step one standard deviation.

And 95 percent population lies within the range of mean plus minus 2 standard deviation and 99.99 percent that is almost all your data should lie within 3 standard deviation plus or minus of your mean or median. So, this is a generally very convenient method of scaling and centering that you can perform comparison of two or more different distributions in terms of z score. So, what are the other methods available for scaling and centric?.

In z score we assume that your data distribution is completely normal distribution or Gaussian distribution; but it is not always the case. In real example or in real life examples we can see that the distribution may not be perfectly normal. And in that case assuming that to be a normal distribution is not the right way to go on. So, in that case we can take up other scaling methods such as we can scale our data from a 0 to 1 scales; this is particularly useful, when variables come from possibly different distributions.

But it preserves the shape of each distribution, while making them easily comparable on the same scale. So, here if we call the value as a score you can see that x minus minimum upon maximum. So, each of your observation is subtracted with the minimum value of all observations and then divided by the maximum value of all observations. This brings your data to the range of 0 to 1 and no other data will lie above that range or below that range.

So, in this way we can compare different distribution in terms of scaling and centering.

(Refer Slide Time: 14:33)

Missing values

Sources:

1. The peptide is present in some sample but missing in others
2. The peptide is present below detection limits.
3. The peptide is detected but the abundance is reported as NA due to technical limitations.
4. The peptide might have been present in very low abundance and out-competed by other ions in getting detected.

Categories:

- Missing completely at random (MCAR)
- Abundance dependent missing values



IIT Bombay

Now, next our point is the missing values. As we have already told that missing values are a common feature in omics data. Now let us think about what could be the source of missing values. The first reason could be that peptide is present in some samples and absent in some samples. That is it is always generally a case that if you take 10 control samples or 10 disease samples or cancer samples out of 10 only you could get in 6 or 7 samples that particular protein.

And the rest of the 4 samples if this protein is not present and this can happen this can happen due to various reasons. The reasons could include the peptide may be present in below detection limit in the sample or the peptide is detected; but abundance is reported as NA not applicable due to technical limitations. If in data analysis step missing values can be imputed and there are many other methods for imputation of missing values.

And missing value imputation should ideally rely on the reason behind getting a missing value. Now, next we talk about the categories of missing value. The categories of missing value is based on two properties; missing completely at random that is MCAR abundance dependent missing values. Missing completely at random values occur due to some technical glitches; in instrumentation such as poor ionization, other peptides competing for charges etcetera.

And abundance dependent missing values occur due to peptides below detection limit or even they may not be present in a sample or it may be the case the detector got saturated and fails to record the abundance. So, these are the main two types of categories of missing values. Now, coming to the imputation of missing values here we will be talking about imputation using mean or median or lowest values. So, the mean can be used for each protein across the samples in a particular control or disease group, we can use the mean to replace the missing values there.

(Refer Slide Time: 16:41)

Missing value imputation

	Control 1	Control 2	Control 3	Control 4	Diseased 1	Diseased 2	Diseased 3	Diseased 4		Control 1	Control 2	Control 3	Control 4	Diseased 1	Diseased 2	Diseased 3	Diseased 4	
Gene A	23	21	25	28	35	36				Gene A	23	21	25	28	35	36	35.5	35.5
Gene B	25	22	24		34		34	39		Gene B	25	22	24	24	34	36.5	34	39
Gene C		22			29			27	26	Gene C	21	22	21	21	29	27.33	27	26
Gene D	23		19		31				34	Gene D	23	21	19	21	31	32.5	32.5	34
Gene E	25	23		22				33	35	Gene E	25	23	23	22	34	34	33	35
Gene F		25	27	29		26	24			Gene F	27	25	27	29	25	26	24	25
Gene G		27	29		27		32			Gene G	28	27	29	28	27	26	32	31
Gene H	21		23			33			40	Gene H	21	22	23	22	26	33	26	40
Gene I	31	34			31	31			33	Gene I	31	34	21	21	31	31	26	33
Gene J			32	25	21	35		28	29	Gene J	21	32	25	21	35	28	28	29

Red: replacement by means

Green: replacement by lowest observed value

Blue: replacement by median

Control

Diseased



IIT Bombay

Or we can use the median values to replace the missing values. However, this method has some pitfalls because, this method may not truly reflect the biological variations; since we are substituting the multiple entities with the same values. So, it is not a true reflection of biological variations. Now, in this slide you can look at the example of missing value imputation using mean, median, or lowest observed values. As you can see in the matrix here that we have genes on the row side and column side.

And control samples that is divided into controls and diseased on the row side. So, here 4 control and 4 disease samples, you can see this kind of values are a lot of there are a lot of missing values. For gene A, this is missing for disease 3 and diseased 4 sample, gene B for this is missing for control 4 sample and so on. So, this kind of missing values always

interferes with further data analysis. And based on the distribution and based on some logic we can impute the missing values so, that we get a complete matrix for further analysis.

As you can see the on the right hand side, there is replacement values by means; that is the red ones are replaced by the value of the mean. If you see at the diseased category on the right hand side the 35 and 36 are the values for diseased 1 and diseased 2 sample. But the disease 3 and diseased 4 samples are missing these values can be imputed with the mean that is the first two diseased values the mean value is 35.5. And we can impute with the mean values.

So, this is a method of imputation for using mean values. Again if you look at the replacement by medium that is, highlighted in blue colour that you can see. For gene E that the control samples are having a median of 23 in 23 25 and 22; these at the 3 ranges of the control 3, 4 control samples. So, we can impute the missing value as by median using 23 for the gene e in control 3 sample.

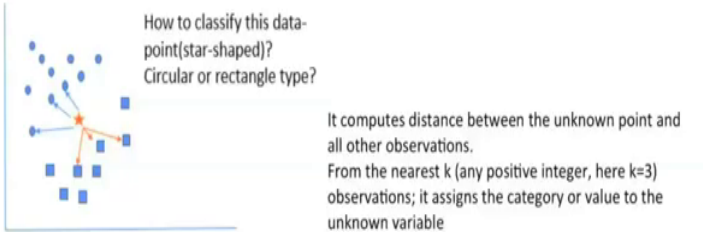
So, this is another method of this is the method of median based missing value imputation. And the last one green highlights the replacement by lowest observed value. So, it assumes that the peptide is below the detection limit in the sample; that is why we are encouraged to use this kind of missing value imputation by the lowest observed value and this is shown in green. In green colour as you can see for the control samples the lowest observed value is 21 in all the for all the genes.

So, we can ideally substitute the 21 values, in all the missing places with this kind of missing value imputation procedure. So, this is you got a general summary of how the missing value what the missing value is and how can they be imputed. Now coming to another missing value imputation approach that is k NN.

(Refer Slide Time: 20:00)


kNN (k-Nearest Neighbor) Imputation

- kNN uses an unsupervised clustering algorithm to impute missing values



How to classify this data-point (star-shaped)?
Circular or rectangle type?

It computes distance between the unknown point and all other observations.
From the nearest k (any positive integer, here k=3) observations; it assigns the category or value to the unknown variable



IIT Bombay

That stands for k nearest neighbor and k NN uses and sup unsupervised algorithm to impute the missing values. Now, if you look at the graph that the start shaped point now we need to classify we do not know what which category it is going to be. So, we need to classify this star shaped point whether it goes to the circular type or goes to the rectangular type. Based on this the k NN algorithm what it does it computes the distance between all the unknown points of all the observations. And from the all the observation it takes the nearest k that is any positive integer it should be any positive integer.

And here in this example we are showing that it is k equal to 3; that means, it is taking the 3 shortest distance of the circular type and the rectangular type. And from these distances the lowest distance, it has the sum of lowest distance it will be categorized as the particular category of the unknown variable.

In this case you can see that the it this start point might go to the rectangular side. So, this algorithm works like this. And based on this if it is a rectangular type or a circular type it may be assigned a value based on the total distribution value of that particular category. And this is the underlying assumption of k NN based missing value imputation.



(Refer Slide Time: 21:33)

Normalization

- Normalization is a pre-processing step in omics data analysis; in order to compare different samples to make valid inferences.
- It is performed to remove technical artifacts.

Various normalization methods:

1. Median-MAD (Median Absolute Deviation) normalization
2. Quantile normalization
3. Median normalization



Now, the next topic comes to the normalization. Normalization is a pre processing step in omics data set, in order to compare different samples to make valid inferences. Normalization is performed to remove the technical artifacts in experiments technical artifacts may come from changes in column, may come from manual handling or may come from changes in daily day to day temperature also.

So, this kind of technical artifacts are different from the true biological artifacts. So, the whole aim of the normalization process is to remove these technical artifacts and without

disturbing the biological variations. So, from these technical or biological variations we need to classify whether this data set could be used for comparison. And to remove these technical artifacts, we use various normalization methods such as quantile normalization and median MAD normalization. Median MAD means Median Absolute Devolution normalization.

(Refer Slide Time: 22:35)

Quantile normalization: step by step

- Quantile normalization is performed to make two or more distributions identical in statistical properties.

Genes	Case1	Case2	Case3
A	5	4	3
B	2	1	4
C	3	4	6
D	4	2	8

→

Genes	Case1	Case2	Case3
A	4	3	1
B	1	1	2
C	2	3	3
D	3	2	4

→

Genes	Case1	Case2	Case3
A	2	1	3
B	3	2	4
C	4	4	6
D	5	4	8

→

Genes	Case1	Case2	Case3
A	5.66	4.66	2
B	2	2	3
C	3	4.66	4.66
D	4.66	3	5.66

Raw data

Ranking the values
lower to higher

Sorting according
to rank

Replacing by mean

Pitfalls: Extreme values could be masked; hence may lose potential biomarkers.
Quantile normalization is used generally for microarray data analysis.

IIT Bombay

Now, here in this slide, we have shown you an example of quantile normalization. And quantile normalization is a technique for making two or more distributions identical in statistical properties. So, let us take a case of 4 genes with 3 cases the values are filled arbitrarily and, here I am going to take you through the steps in quantile normalization step by step. So, from the raw data the first task is to rank the values from lower to higher. As you can see in the ranking of the values of lower to higher.

In case 1 the value 5 is given a rank 4 that is it is taking a rank from lower to higher order; the next step is to sort the data according to the rank. So, we will rank the we will sort the values according to the rank that is 1, 2, 3, 4 there are 4 genes here. And after that we will replace the values by means of that particular rank. So, in terms of if you see at the raw data and the final replaced by mean data, that the rank remains the same; the order of genes A, B, C, D remains the same.

But the values are changed and they are having the similar identical statistical properties. So, this is the underlying assumption of quantile normalization. However, there are some pitfalls in quantile normalization that extreme values could be masked. Hence we may lose potential biomarkers; because biomarkers are the proteins of interest that is having dysregulated expression level that is in the normal distribution curve they might be in the tail region.

So, quantile normalization may mask your potential biomarkers from the analysis. And so, it should be utilized or implemented with proper care. And quantile normalization is generally used for microarray data analysis.

(Refer Slide Time: 24:41)

Median-Median Absolute Deviation (MAD) normalization

Median-MAD normalization is preferred when the data distribution is non-normal.

Median Absolute Deviation = $\text{Median}(|X_i - \text{median}(X_i)|)$

MAD scaling is similar to z-score scaling; the only difference is that we take median of all absolute deviations so that they are not affected by extreme values; when the data distribution is non-normal.



IIT Bombay

Next comes the median absolute deviation normalization; what we call in abbreviation as MAD normalization. So, the MAD normalization is preferred when the distribution of the data is not normal and it is particularly useful in that case. z score is particularly applicable when we assume that the data distribution is quite glycine and normally distributed.

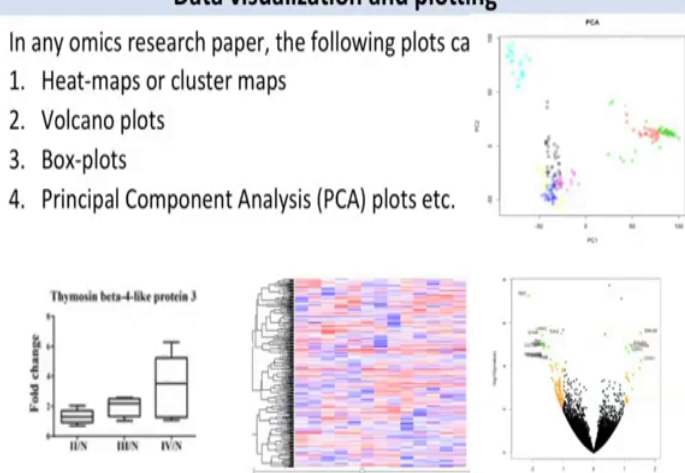
So, MAD normalization is similar to z score scaling and the only difference is that, it takes the median of absolute deviations. So, that they are not affected by extreme values and this is particularly useful for non normal distributions. The median absolute deviation, as you can see in the formula is the median of the absolute deviations.

(Refer Slide Time: 25:32)


Data visualization and plotting

In any omics research paper, the following plots can be used:

1. Heat-maps or cluster maps
2. Volcano plots
3. Box-plots
4. Principal Component Analysis (PCA) plots etc.

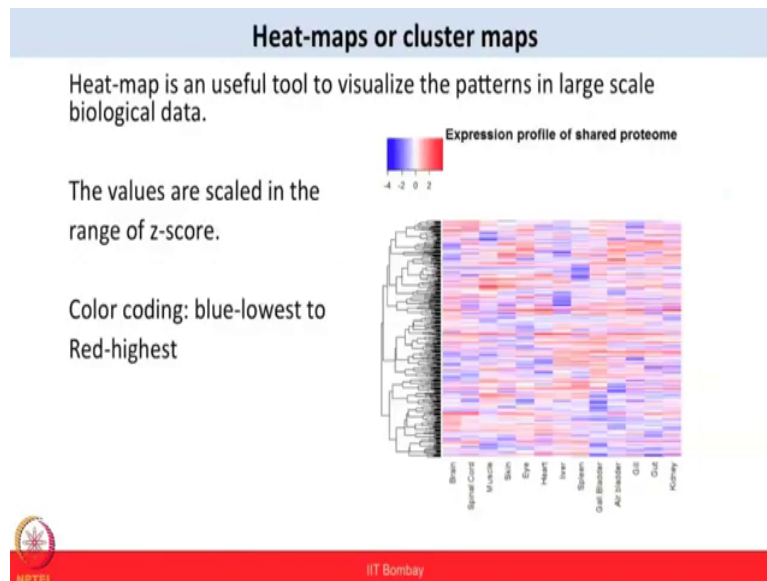


The figure displays four distinct data visualization techniques. At the top right is a PCA plot with axes labeled PC1 and PC2, showing clusters of data points in various colors. Below it on the left is a box-plot titled 'Thymosin beta-4-like protein 3' showing 'Fold change' for three conditions: I/N, II/N, and IV/N. In the center is a heatmap with a dendrogram on the left, showing a color-coded matrix of data points. To the right of the heatmap is a volcano plot with 'log2(Fold Change)' on the x-axis and 'log10(P-value)' on the y-axis, showing a distribution of points with several labeled points.

 IIT Bombay

Now, let us talk about the data visualization and plotting techniques. If you look at any omics paper in proteomics or genomics these kinds of plots are very common; that is heat maps, cluster maps, box plotting, volcano plots, PCA plots so on. So, let me take you through the step through the glimpse of this all data visualization and plotting techniques.

(Refer Slide Time: 26:06)

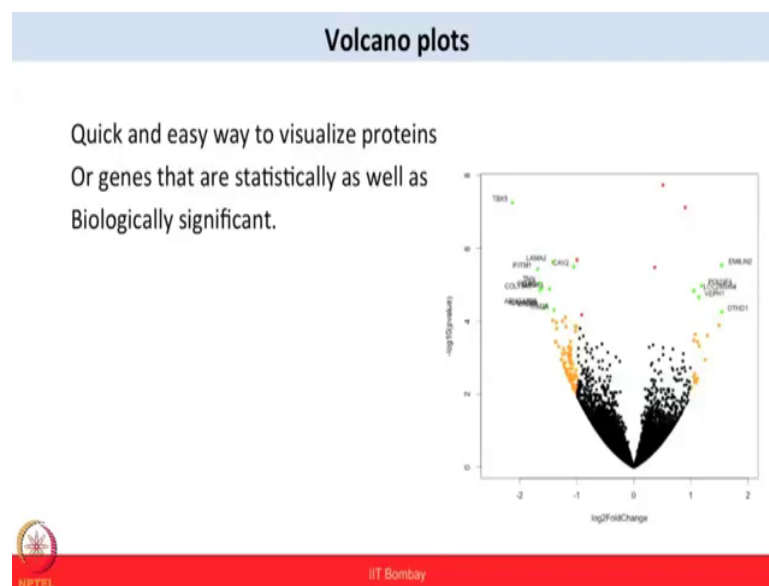


Let us talk first about the heat maps or cluster maps. Heat map is an useful tool to visualize the patterns in large scale biological data. So, what is basically is that there you can actually visualize thousands of genes and their pattern across numerous samples. As you can see in the figure that the rows are the list of proteins and on the column there are list of samples.

And you can see for a particular protein if you go through, but go through the row, you can see the pattern of this particular protein across different samples. And this allows us to get a visualize an overview of all the patterns in the large scale biological data. And this colored scale that ranges from minus 4 to plus 4 this comes from the z score scaling. So, after you log transform your data you do the z score scaling and you can plot this as a heat map; the blue ones are representing the lower expressing values.

And the red ones are expressing the highly express proteins. So, this is a very useful technique in omics studies that you can make patterns out of this whole large scale biological data. Next we talk about the volcano plots. Volcano plots are a quick and easy way to visualize the genes or proteins that are statistically significant.

(Refer Slide Time: 27:36)



In volcano plots what it plot on the x axis there is log two fold change values and on the y axis there is negative logarithm of p values. So, after you perform your T test or ANOVA, you end up with some T values for a particular protein. And you can also determine the fold changes in comparison to the control samples for the disease samples.

So, this disease samples and control samples ratio is called the fold change in terms of expression values. So, if we plot the log 2 of fold change and the negative logarithm of the p values. We can know what are the points that are very much statistically significant in terms

of p values and, their significance in terms of fold change. As you can see in the figure the red dots are above the fold change values.

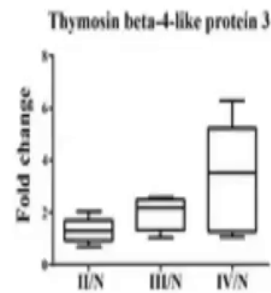
So, these are having significantly significant p values. And the green dots are on the lower on side are having a low 2 fold change in the range of minus 1.5 to minus 2; that is these proteins are quite downregulated as compared to the reference samples and on the right hand side. The green dots denotes this proteins are upregulated as compared to the reference samples. So, this is a quick and easy way to visualize your data and which proteins are of your interest in terms of biological questions.

Now, coming to the box plots, box plots an easy way to visualize the spread of your data points for a particular gene or protein. So, when you have narrowed down to your proteins or genes of interest, this box plot can be an easy way to quickly compare the distribution of the abundance values across the samples.

(Refer Slide Time: 29:35)

Boxplots

Boxplots are used to visualize spread of the data-points for a gene or protein of interest among different subgroups.



Reference:

<https://www.ncbi.nlm.nih.gov/pubmed/28481733>



IIT Bombay

As you can see in the example on the right hand side, that we have taken the thymosin beta 4 like protein 3 ; this protein we have plotted the fold change values across different types of samples. And this is compared in great 2 samples, great 3 samples and great 4 samples.

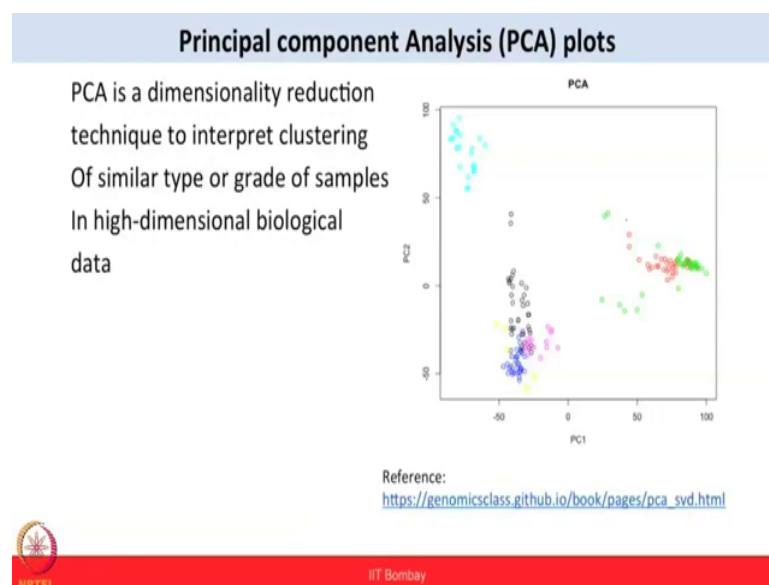
You can see the data distribution of this protein in terms of fold change values across those samples. And this is generally plotted in terms of box plot. And in box plot you can see that there is a median line in the middle of the box and this represents the median value of expression of that particular protein. And after that you have on the upper hand side, you have third quantile and on the lower hand side you have first quantile of the expression values.

So, this is are the parameters of box plots that is used to actually visualize the spread of the whole data. Now, let us talk about PCA or Principal Component Analysis; principal component analysis is a dimensionality reduction technique to impar interpret the clustering

of similar type or grades of samples. In PCA generally what you see is a 2 dimensional plot the x axis represents the principal component 1 and the y axis represents the principal component 2.

So, the principal component 1 represents the maximum variation of your sample. So, this is the line where you can group or cluster your data and see how they are clustering.

(Refer Slide Time: 31:17)



As you can see there are different colors of grouping of individual samples into different clusters; that means, they are quite different from the other samples. This PCA plot is another easy way to visualize the grouping and clustering in your samples for making further biological inferences.

(Refer Slide Time: 31:38)

Points to Ponder

- Missing values in the datasets can be imputed using different replacement methods like mean replacement, median replacement or by using k-Nearest Neighbour (kNN) method
- Log transformation and scaling of data is performed to reduce the skewness of the data and to make data more interpretable and inferential
- Normalization, an important step in biological dataset, is performed to remove the day-to-day variation and technical variability to make sure the results obtained are due to actual biological differences
- Further, statistical analysis is performed, to select the list of significantly altered biomolecules
- Owing to the big dataset, that omics studies offer, data visualization and plotting helps in finding the signature patterns providing an overview of the results attained



MOCC-NPTEL
NPTEL

IIT Bombay

I hope today's session was informative. And you got to learn many new tools, which you should start playing with it. There are a lot of data sets available in the public repositories. Of course, you should ask even our lab to provide you some more raw dataset which you can play with you can try to analyse yourself. And with these sessions I hope you are getting good understanding of different steps involved in data processing and analysis. Of course, each instrument platform whether we talk about microarrays or surface plasmon resonance or next generation sequencer or mass spectrometer the initial raw data processing is very unique for each type of instrument platform.

But once you have obtained the data in the excel sheet format, then lot of things are pretty much common; because you want to make biological sense out of these raw numbers and these you know big data tables. And what you are looking at what was your background, what is the noise, what is the actual signal. So, many things for example, you know the processing

steps remains very similar irrespective of which technology platform we are using; of course, there can be uniqueness looking at the constraints of each technology as well.

But finally, what you want to you want to make a biological sense of the data and you want to plot the data in more meaningful way. So, different ways of visualization of data and tools available, which can effectively represent the data are very crucial. So, some of these tools, I hope what you learn today would have given you much better understanding that is how effectively you can start looking at your data.

And in the upcoming lecture, we will continue further on this and show you more software and analysis, which can now try to build from your protein and gene list how to we can make the networks and pathways. And make better biological sense looking at the complexity of the problem, which you are working on; how to make sense of that information by targeting different pathways and interaction networks.

We will continue more of the session in the next lecture till then.

Thank you.