An Introduction to Proteogenomics
Dr. Sanjeeva Srivastava
Dr. D. R. Mani
Department of Biosciences and Bioengineering
Principal Computational Scientist
Indian Institute of Technology, Bombay
Broad Institute of MIT and Harvard, USA

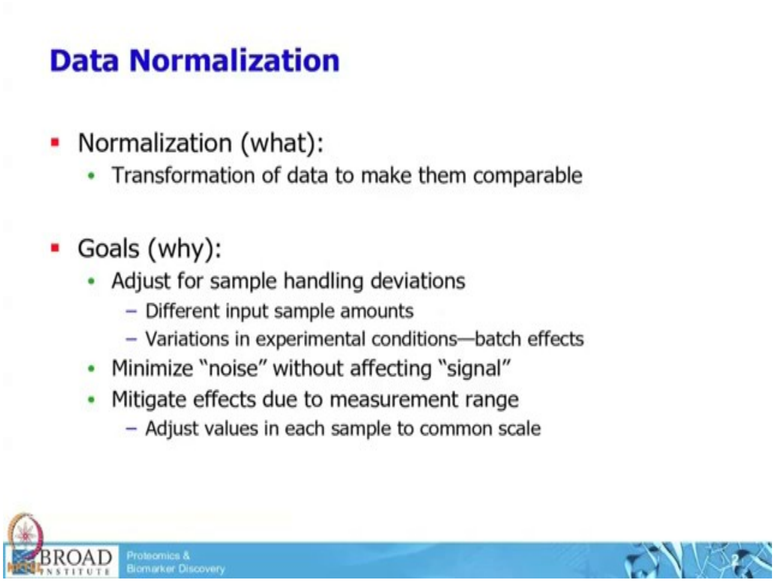
Lecture - 18
Data Analysis: Normalization

Welcome to MOOC course on Introduction to Proteogenomics. In the previous lectures you have seen how new technologies like genomics and proteomics could generate big datasets. However, to obtain the meaningful insights from these data, you need to utilize various statistical and computational tools.

In this slide, today Dr. D. R. Mani from Broad Institute; is going to give you the first lecture and he is going to talk to you about Data Normalization, which is a very important aspect of omics data analysis. There is lot of variations during your manual sample preparation steps as well as the artifacts which might be coming because of the running issues or the instrument, day to day variability or the samples batch to batch variability; how to rectify some of this information and correct for these variability is very crucial.

In this slide, normalization techniques allow simultaneous correction of the various issues which one could see because of the instrumentation; such as in the mass spectrometry context ionization efficiency of the detected peak or even achieving more quantitative values can be better obtained after normalization. So, today Dr. Mani is going to talk to you about data normalization which is very important aspect of omics data analysis. He will explain to you different strategies which could be employed for doing normalization like quantile normalization, median normalization, median MAD normalization and many other methods. So, let us welcome Dr. Mani for his lecture on data normalization.

(Refer Slide Time: 02:17)



Data Normalization

- Normalization (what):
  - Transformation of data to make them comparable
- Goals (why):
  - Adjust for sample handling deviations
    - Different input sample amounts
    - Variations in experimental conditions—batch effects
  - Minimize "noise" without affecting "signal"
  - Mitigate effects due to measurement range
    - Adjust values in each sample to common scale

BROAD INSTITUTE
Proteomics & Biomarker Discovery

Let us start with the data normalization. So, what is normalization? Normalization is transforming data, so that they are compatible. So, you have 20 TMT reactions you ran and so you have 20 ten plexes, you got data from all those; now you want to compare them all together. For whatever reasons each one is slightly different, so you want to put them on the same scale so you can compare them. So, that is the purpose of normalization.

So, what do you hope to accomplish by doing it? So, you can adjust for sample handling deviations. So, in one sample when you pipette at the sample you got slightly more and the other one you got slightly less or the temperature on one day was different from another and so you got something that was slightly different those kind of slight sample to sample differences can be taken care of.

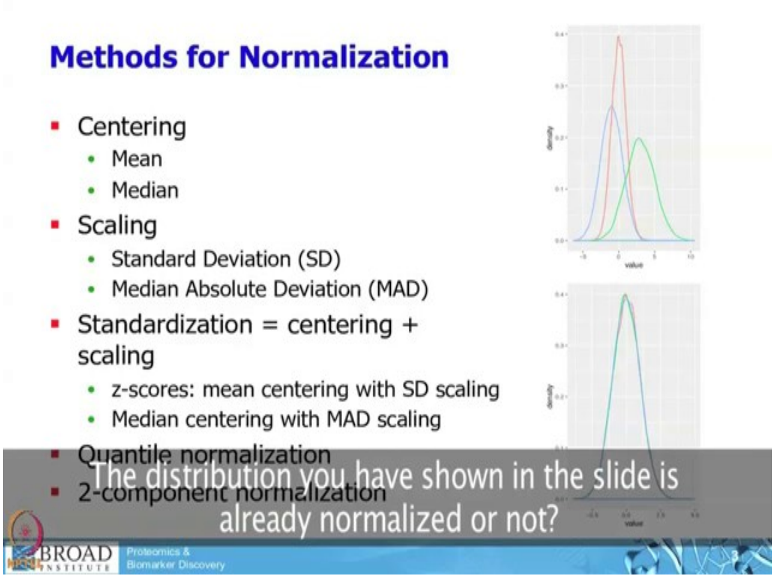
You could also have slight differences in experimental conditions. So, those are batch effects I think I will get to those separately, but if the batch effect is very small and kind of just sample specific; you could be able to get rid of it using normalization. The whole purpose of that is to you have peptides or proteins that you are interested in that are you that is your signal.

So, things that are different between your different types of cancer or between cancer and control that you are interested in. So, things that are really different is called signal because that is what you are looking for, but then because of the measurement process and because of technical variation and biological way because of primarily technical; variation there is a lot of noise that is introduced in your data, the purpose of normalization is to minimize noise and maximize signal.

So, in I think I have slides about this later, but this is a point that is worth repeating. So, I said the purpose of normalization is to minimize noise and maximize signal. So, that kind of assumes that you know what noise is and what signal is. In most real world experiments you do not know where the line that divides noise and signal is. So, in trying to get rid of noise; if you overdo your correction, you will get rid of signal and if you under correct you will have a lot more noise than you want.

So, this is a; very kind of process that you have to do very carefully and be thoughtful about what you are doing and how it affects the questions you are going to ask about the experiment. So, there is usually no one procedure that works all the time. So, you have to carefully look at what your experiment is, what your questions that you want to get answered from the experimenters and then make sure that the normalization process is appropriate for that.

(Refer Slide Time: 05:03)



Methods for Normalization

- Centering
  - Mean
  - Median
- Scaling
  - Standard Deviation (SD)
  - Median Absolute Deviation (MAD)
- Standardization = centering + scaling
  - z-scores: mean centering with SD scaling
  - Median centering with MAD scaling
- Quantile normalization
- 2-component normalization

The distribution you have shown in the slide is already normalized or not?

So, there are many ways of doing normalization, the simple ones are called centering. So, basically you start; so what I am showing here is so again remember we are talking of log ratios. So, this is a histogram of log ratio. So, how many ratios have value 0? So, it is that many, how many have value 1; it is that many and so on and each color is the sample. So, just for a mockup; I am showing three samples and I am showing the distribution of log ratios for each of those samples. So, you can see that the green one is centered around 0.5, the red one is centered around 0, the blue one is centered around minus 0.5 approximately. So; that means, the average ratio was slightly off which means the amount you put in was slightly off.

So, one thing you can do is you can say I want all my peaks lined up and that is centering. So, you subtract the mean from each one of these or the median; they will all line up and that is called centering, if you do the mean you will be affected more by extreme values or outliers. So, your actual biomarkers will have an effect on how much you move which you probably do not want, in which case you would use the median which is much more robust. And so it will kind of ignore outliers and use the central part of the distribution to decide how much to move.

And then the second thing is scaling. So, you can see that the green one is much more fatter than the red one; the red one is a thin long peak whereas, the green one is a spread out peak. So, in order to make things more comparable, it would be helpful in many situations to have the spread also be equal. So, in other words you want to scale it so that the standard deviation or the fatness of your distribution is essentially the same. And so to do that you divide a by the standard deviation or there is a more robust measure called median absolute deviation, which is short for the MAD is something that short for that.

So, when you divide by one of these values for that distribution; then the spread also becomes the same. And so if you do both centering and scaling it is called standardization. So, in standardization you first center by subtracting the mean or median and then scale by dividing; by the standard deviation or median absolute deviation. So, when you do that if you take these three samples in the top and you do standardization, you get what you see at the bottom. So, you can see now they are all lined up, their samples are lined up, their spread is the same and so the samples are now normalized and you can look at them.

Student: Scaling is not the same you see it do not move down.

Which scaling is not the same?

Student: The upper one, I mean you have the three peaks.

Yeah.

Student: After centering and scaling.

So, the upper one is unnormalized data; after you do centering and scaling you get what is in the bottom.

Student: Yeah.

And you are saying the bottom is not lining up?

Student: Yeah.

Well, it may not be exact; so that is the thing right; so not all your samples are exactly the same and your standard deviation and so the distributions also may not be exactly identical; in terms of the shape of the distribution. So, if you had all theoretical perfectly distributed normal Gaussian distributions, then they would all line up. But these are real data with kinks in the middle and not exactly normally distributed and we are doing transformation that kind of assumes they are sort of normal.

Student: Some of those outliers are very informative other than the red

So, the only outliers that would kind of affective are the one said that extreme tails because those are your markers because those are the ones that are most different in your sample right. So, if the ratio is like log 2 ratio is 2, then; that means, that protein was four fold upregulated compared to your reference; whereas, if you are in the negative side then that protein was significantly down regulated compared to your reference.

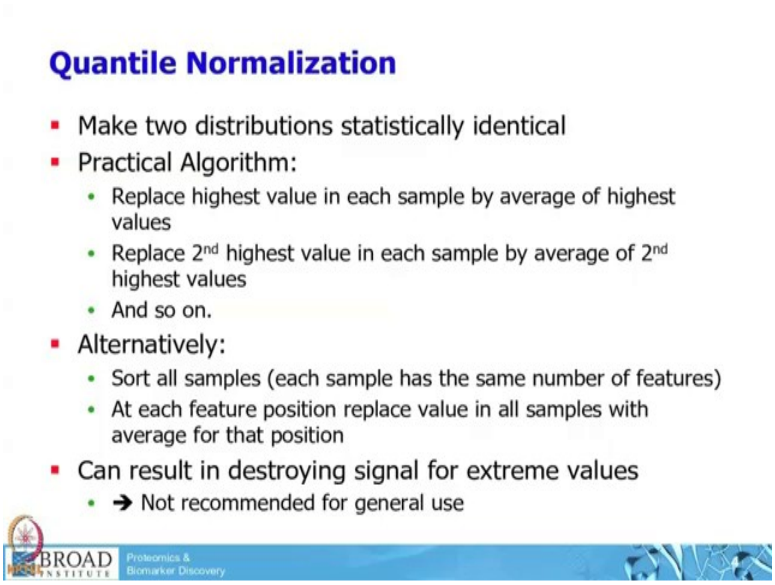
So, the reference is kind of the average; so something that is very different from the averages what is going to affect your analysis. So, if there are little mismatches in the middle; I think it will not be too much of a problem, but what you need to be careful about are the tails. And actually we will talk a lot more about tails when we go to the other two types of normalization that is listed there.

So, in terms of terms to use when you say when people say I, z scored my data; z scoring means you did standardization by mean centering and standard deviation scaling; so z scoring is a term for that. And if you use the median version you median center and scale by MAD; it is called median MADs normalization or median MAD centering; no, median centering with MAD scale; there is no special name for this people just describe it.

So, there are two other types of normalization I have listed here one is called quantile normalization and the other is called two component normalization. So, we just had a question and we discussed that what matters really in these distributions are the tails of the distribution. So, you want the all the proteins that are of interest to you or in the tail.

So, these are the most up regulated proteins in that; in your set of samples and on the other side or the most down regulated proteins in the set of samples. And so those are the things you are going to be interested in. And you want to make sure your normalization process does not mess with those to the degree possible and so let us look at these two other procedures.

(Refer Slide Time: 10:55)



Quantile Normalization

- Make two distributions statistically identical
- Practical Algorithm:
  - Replace highest value in each sample by average of highest values
  - Replace 2nd highest value in each sample by average of 2nd highest values
  - And so on.
- Alternatively:
  - Sort all samples (each sample has the same number of features)
  - At each feature position replace value in all samples with average for that position
- Can result in destroying signal for extreme values
  - ➔ Not recommended for general use

So, quantile normalization basically takes two distributions and makes them statistically identical by doing the following. It sorts all the values for all your samples and then it picks the highest value that you have and then says the normalized value is basically the median of the highest value I have seen in all my samples.

So, let us say you had two different types of breast cancer samples in your study. One of them had a protein that was highly up regulated and the other one that protein was not up regulated. Now, in this case the highly up regulated protein is going to be a kind of squash to conform to the median because it was not up regulated in the other one; well that is not completely true because you are sorting just by value.

So, if there one sample did not have too many things happening. So, let us say it was a normal sample; everything was around average and the other was a cancer sample where some proteins were very high and some proteins were very low. Now, when you sort all the values the highest value in the normal sample will be some normal protein; whereas, the highest value in the cancer samples will be this biomarker you are looking for. And then when you do quantile normalization it is going to take all those values and take the median which is going to be somewhere in the middle. So, you just lost your biomarker.

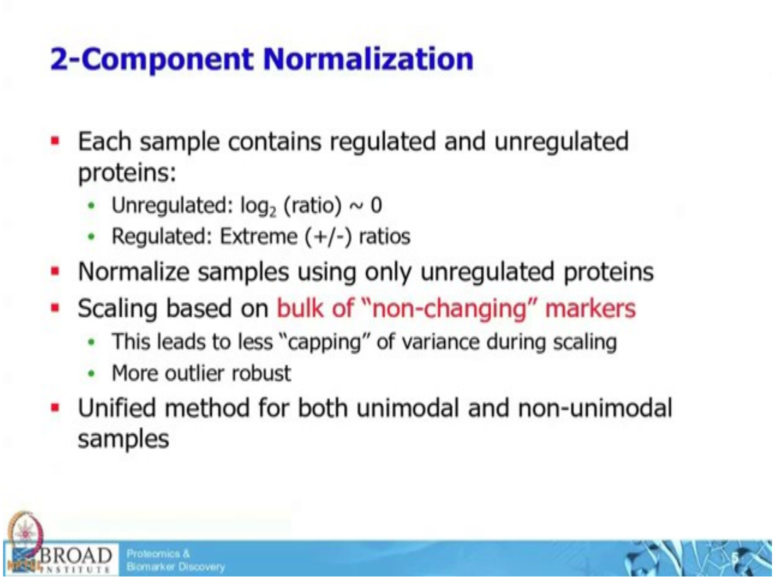
So, if you do quantile normalization it is kind of I, I call it destructive normalization because you lose a your; you there is the possibility that you could lose signal. So, this was introduced to work with affymetrix microarray data and that kind of data, where you have set of numbers and you know the range of numbers is going to be between 0 and 20000, but for some reason in one of the cases the numbers ended up being between 100 and 7000.

So, then you can say I am going to make my 7000 thing closer to 20000 or vice versa, but in real world projects when you use quantile normalization; you have to be really careful that you know exactly what you are doing and you are sure that it is not like destroying signal that you might have in your study.

So, in order to deal with this problems; one way we came up with to address normalization and what happens in the extremes is called two component normalization. So, the concept here is if you look at the distribution of proteins. So, you have a large number of proteins in the middle with an average ratio log ratio of 0. So, those proteins are not changing between your samples and your reference. So, those are kind of the unchanging proteins and then there are proteins in the tails that are either up or down regulated compared to the reference.

So, what you want to do is normalize based only on the unchanging proteins so that you leave the extreme proteins alone; you do not mess with them.

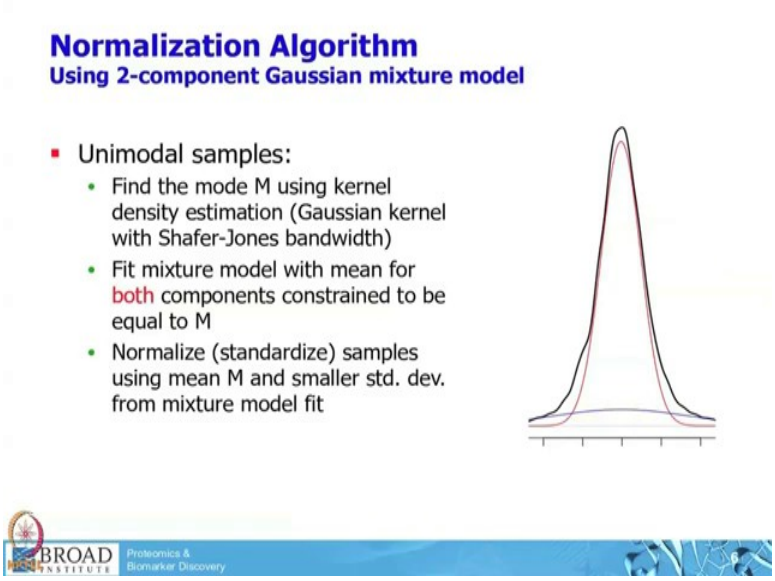
(Refer Slide Time: 13:57)



2-Component Normalization

- Each sample contains regulated and unregulated proteins:
  - Unregulated: $\log_2$ (ratio) ~ 0
  - Regulated: Extreme (+/-) ratios
- Normalize samples using only unregulated proteins
- Scaling based on bulk of "non-changing" markers
  - This leads to less "capping" of variance during scaling
  - More outlier robust
- Unified method for both unimodal and non-unimodal samples

(Refer Slide Time: 14:01)



Normalization Algorithm

Using 2-component Gaussian mixture model

- Unimodal samples:
  - Find the mode M using kernel density estimation (Gaussian kernel with Shafer-Jones bandwidth)
  - Fit mixture model with mean for both components constrained to be equal to M
  - Normalize (standardize) samples using mean M and smaller std. dev. from mixture model fit

So, to do that what we do is; there is a mechanism called a mixture model; where you can say I know there are two different distributions mixed up in my the plot that I show. So, in other words I have a plot like this which is the black.

So, the black has things that are not changing in the middle and things that are changing in the tails. I want to just find out those set of proteins in the middle that are not changing. So; that means, there are two distributions in my black line; one is the thing that represents proteins that are not changing and another is a distribution that represent things that are changing.

And so we fit two different Gaussian or normal distributions using a process called mixture modelling. And what that does is it says I mean you told me there are two distributions; let me look at the data and figure out where the two distributions are, what their mean and standard deviation is. So, this procedure will come up with this red curve which says my mean for this is 0 and the standard deviation is like say 1. And then the blue curve which represents the proteins that are changing for which the mean is also 0, but the standard deviation is 100 because they are very spread out.

Then you say the red curve represents the proteins that are not changing. So, I am going to use only the red curve to do my normalization. So, this is basically like z scoring, but you z score with a specific distribution that was calculated to include only the proteins that are not changing. So, this way you end up not messing with the proteins that are changing and kind of focus your normalization only on the unchanging proteins, there is no one normalization that is good for everything.

So, usually what we do is we do two component normalization in most big studies, but if there is. So the two component normalization assumes there is a set of proteins that are not changing. So, suppose you did a protein-protein interaction experiment, where you use some immunoprecipitation for some protein of interest and you pull down that protein and then you did proteomics on what came down. There are most likely a lot of things are changing there is may not be a set of proteins that are not changing. So, in that case you cannot use two component normalization; because it is assuming there is a set of things that are there that are not changing. So, what would you use in that case? You can do median MAD normalization.

So, usually what we do is; we look at two component normalization if it seems appropriate and if it is working, then we will use that. If that is not the case; then we will the default that we fall back to is median MAD normalization, but in some cases even that may not be possible. So, we have had experiments where they had a control sample and a treated sample, the mean or the median for the control sample was very different from that for the treated sample.

So, then if you do this kind of normalization then you would end up with pulling everything together and then you mean that may not be the appropriate thing to do. So, then we split it into two groups and do like normalization for each group. So, it really depends on the data you have and the experiment you have done and what the questions are that you are addressing through the experiment.

So, if there is one thing you always want to try maybe median MAD would be a simple thing to do. So, median MAD normalization; you can probably do it in excel; you do not need any like additional software. But for two component normalization you will need some kind of a statistical analysis package.

So, we do it in R, but there are a lot of others that can implement the same process or in your hands on we are going to look at prodigy which has two component normalization implemented in it; you can explore that on your own and see. I think, we probably will not do that today because it takes a while to do the normalization. So, you might want to try it on your own, but in general this one and median MAD or what we generally look at; unless there is a reason not to.

Student: So, if we are taking that human protein; like plasma protein or something.

Yeah.

Student: Then in that case, if we are if we want to do a two component normalization then what should be the number like minimum number of protein we should look at to be like not changing so that we can take it into consideration because we are having three different clinical conditions. So, in that there will be very less protein, which are not changing across all the three conditions.

Yeah.

Student: So, in that case what should be the minimum number of protein?

So, usually my rule of thumb is; you should expect there to be at least a few hundred proteins that are not changing.

Student: Ok.

So, if you have like only 5 or 10 or very few in the tens; then it is not a good method to apply, if you have a few hundred or beyond that. So, in most like discovery experiments like the CPTAC experiments; we have thousands of proteins that are expected not to change. So, in those situations these this will definitely be applicable, but if you have only 300 proteins and you think the large fraction of them would be different from yours the various experiments; then the I would not use this I would just use median MAD normalization.

Student: Can we use Pareto scaling as well?

Pardon.

Student: Can we use Pareto scaling as well for proteomics data?

Sure, so Pareto scaling is if you expect fatter tails. So, here the two component normalization is kind of this see here. So, it is trying to accommodate for fat tails right. So, what it is doing is you are usually you have a normally distributed; you assume normal distribution, but in most real data the center part is normally distributed, but you have fatter and fatter tails.

So, if you expect a lot of things that are changing; then you can try Pareto distribution, but I think that might still need a central part that is not changing, but it will accommodate tails that are fatter. So, if there are a larger proportion of things that are changing; I think you can deal it deal with that using that distribution.

Student: Here are we not overestimating the fold change compared to like unchanged protein like in earlier we were under estimating it, here we are over estimating.

Yeah.

Student: There will be more than actual.

So, the fold change is only relative to the reference right. So, and you are trying to make it the same that the measure the same for all your samples. So, if the problem you are saying will come when you do this with like actual intensities, but that is kind of why we; we take ratios and log transform.

Student: So, basically you are saying we should normalize the normalized data?

You should not.

Student: Ok.

If you do not like the normalization that the software is doing; then you should disable the normalization and do this or something else separately, but you should not normalize data that is already been normalized. So, there is some other yeah.

Student: I just wanted to clarify that the median MAD that you talked about.

Yeah.

Student: Is for quantitative interact tool right?

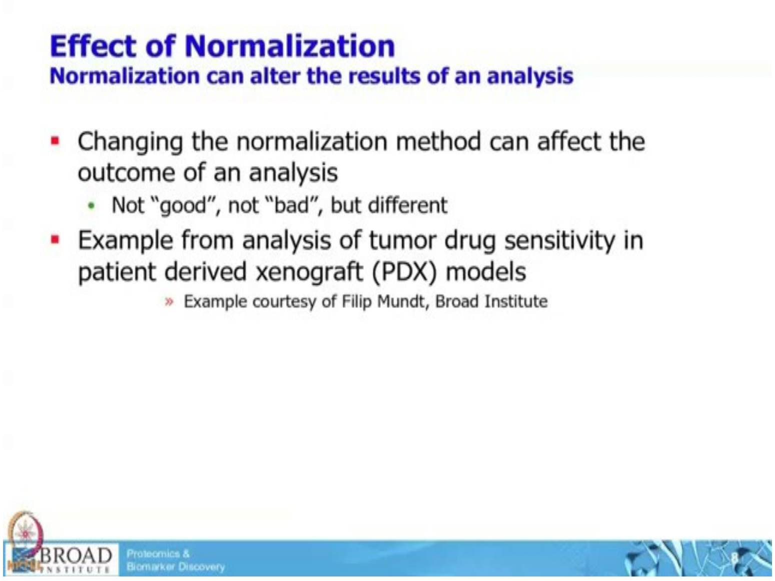
Now, in theory it is for any collection of data, but sure.

Student: Smaller set?

For smaller sets, yeah; if you have smaller sets that is a more reasonable thing to do compared to two component normalization because the number of things that are changing would be smaller. So, changing the normalization method will; obviously, affect the outcome of an analysis.

So, things that were differential markers in one normalization, may not be in another one or may have a lower or higher p value in other words changing the normalization would change the outcome of your analysis. So, that is kind of why you want to think about how you are normalizing; well before you start looking at results, otherwise you will tend to tweak your normalization to get results that are more agreeable which is not the right statistical approach.

(Refer Slide Time: 22:27)



Effect of Normalization

Normalization can alter the results of an analysis

- Changing the normalization method can affect the outcome of an analysis
  - Not "good", not "bad", but different
- Example from analysis of tumor drug sensitivity in patient derived xenograft (PDX) models
  - » Example courtesy of Filip Mundt, Broad Institute

BROAD INSTITUTE Proteomics & Biomarker Discovery

So, like I am saying here the results will be different, but it is not good in one case or bad in another case; it is just different. And so you want to make sure that it agrees with your methodology and experiment and the questions you are addressing before you kind of go ahead.

Student: Sir, I have a question.

Yeah.

Student: Whether the reference is the same as the control?

No, the reference is basically a sample we have to kind of make sure that there is the same thing in every TMT templates so that if there are differences; then we can make this thing the same to kind of normalize things across and.

Student: like the quality reference.

Yes, if you had a control or like a normal sample that would also be compared to the reference. So, as you can see the advantage of this kind of an approach is that you are taking relative ratios to the reference for every sample you have in your channels. And so because of that most batch effects that you might have are kind of taken care of right there. So, you do not need to do like batch correction and other kinds of manipulation of the data the downside is that you are looking at only relative ratios to the reference. So, if you wanted to know is one protein higher in a sample compared to another protein, you cannot answer that question because all you have are relative ratios for protein A and relative ratios for protein B; so you cannot compare across proteins.

But if you want to know whether a protein was high in sample 1 versus sample 2; you can because we are measuring relative ratios to the reference. So, the advantage here is that it minimizes the manipulation of data you need to do later and kind of takes care of batch effects and other kind of systematic technical artifacts; in a more agreeable way, but the downside is that you cannot make like absolute level comparisons.

Student: Sir, if I am working in ovarian cancer.

Yeah.

Student: Reference sample is like it has to be ovarian tissue or could be?

Preferably, because there could be; there could be proteins that are specific to ovarian cancer that may not be occurring in other types of cancers or other normal tissue that you could include and if you do not have it in the reference; you are more likely to miss it.

Student: Sir, if we do not have a normal control.

Yeah.

Student: Can we pool the mixture of that tumours and use as an internal control?

That is kind of what we are doing here. So, we are calling that the reference because we are using it to take relative ratios, but this. So, in the breast cancer project that you are going to hear about today; the CPTAC prospective in a retrospective breast cancer project; the control the reference was basically a collection of tumours. So, we had like 100 tumours there, I think 40 of those went into the reference, but we made sure that the proportion of the different types was the same.

Student: Sir, how do you ensure that the results will not change from lab to lab and its consistent?

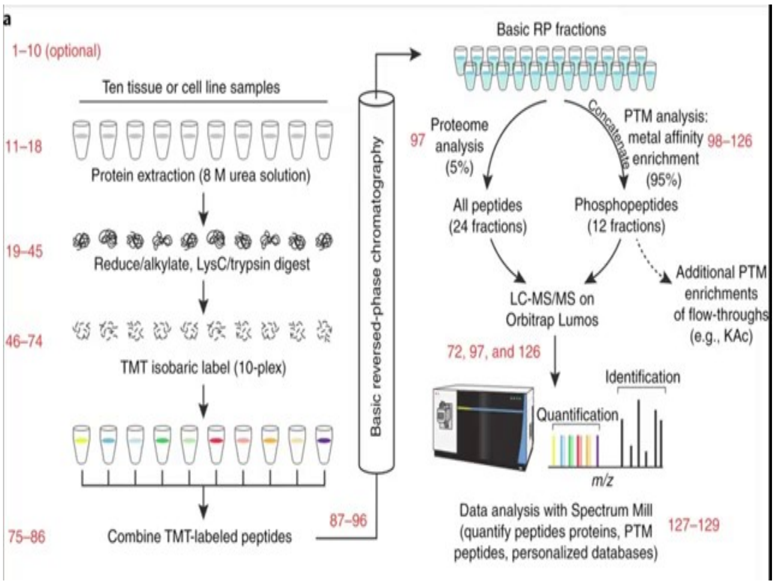
Yeah, so you first off you start off with the protocols like a standard operating protocol that everybody uses. I think the question was how do you ensure that things do not change from lab to lab and its consistent? So, there have been there is been a lot of effort in the CPTAC to make sure that things are reproducible within a lab and across labs. So, actually a lot of the work you are seeing here is what we call CPTAC 2; which is like the breast ovarian and colorectal cancer.

And then CPTAC 3 is we have mentioned some of it like the lung cancer and the newer samples, but CPTAC 1 was basically a 3 or 5 year project; whose entire goal was to make sure that proteomics is reproducible across labs. So, you need to setup your proper SOP's, have similar like have a common sample you run to see if you are getting the same results. So, you just have to go through the process and make sure that things are reproducible.

Student: Whether any standardization available to other lab for standardize protocol?

Right now, I think the answer is no. I think the NCI is trying to come up with something like that, but right now there is not anything that you can get from some lab or some institution to share across labs yeah.

(Refer Slide Time: 27:17)


a
1–10 (optional)
Ten tissue or cell line samples
11–18
Protein extraction (8 M urea solution)
19–45
Reduce/alkylate, LysC/trypsin digest
46–74
TMT isobaric label (10-plex)
75–86
Combine TMT-labeled peptides
87–96
Basic reversed-phase chromatography
Basic RP fractions
Concatenate
97
Proteome analysis (5%)
PTM analysis: metal affinity enrichment (95%)
98–126
All peptides (24 fractions)
Phosphopeptides (12 fractions)
Additional PTM enrichments of flow-throughs (e.g., KAc)
LC-MS/MS on Orbitrap Lumos
72, 97, and 126
Identification
Quantification
m/z
Data analysis with Spectrum Mill (quantify peptides proteins, PTM peptides, personalized databases)
127–129

So, in your slides there is a reference here to a paper in nature protocols. So, that was like a standardized protocol for TMT 10; that was the kind of derived from all the CPTAC labs and has been published. So, if you want like a standard protocol for TMT; so that is a good place to look at.

So, the question is if you start fresh on a new project and you do not have any references how do you do it. So, I think the way I have presented here you use the; you create a reference for each big project. So, if you have a project where you how only like 4 samples you are running in duplicate that is the TMT templates, you use 8 channels and maybe a couple of replicates and that is it; so you do not need a reference.

But if you have a project where you have a 100 samples; then you have to first get enough of those samples in, create your reference by combining the equal amounts of multiple samples and then start. So, for all these projects we did not have any reference before we started. For the breast cancer project, we started by creating the reference, for the lung cancer we started by creating a new reference. So, you have to start by creating a reference.

Student: So, in a breast cancer you have the several types of breast cancer.

Yeah.

Student: So, you have the reference for each of I mean.

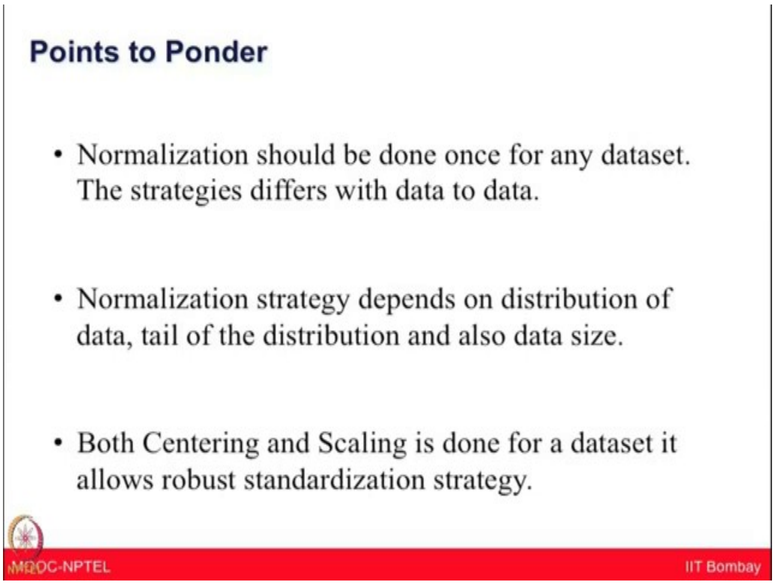
No, but in the reference you have the same fraction of each of those as you have in your samples.

Student: If I say fraction means you combine in its different types of breast cancer.

Yes yeah. So, if you had like let us say you had 25 percent each of four different types of cancer, four different subtypes in breast cancer; then when you create your reference you want one fourth of each of the different types in your. So, let us say you have 100 samples, but you have enough material to create the reference from only 60 of the samples, then you want 15 from type 1, 15 from type 2 and so on.

So, the proportion of the different subtypes that go into your reference should match the actual proportions in your sample set, so that it you are not overly emphasizing one or other group. It is just a guideline, if it is off by a little bit it should be fine, if it is completely off or you left out one or two subtypes; then it is more likely that proteins that are specific to that subtype will not be observed which is a disadvantage. So, your perfect biomarker for that subtype will not be there.

(Refer Slide Time: 29:57)

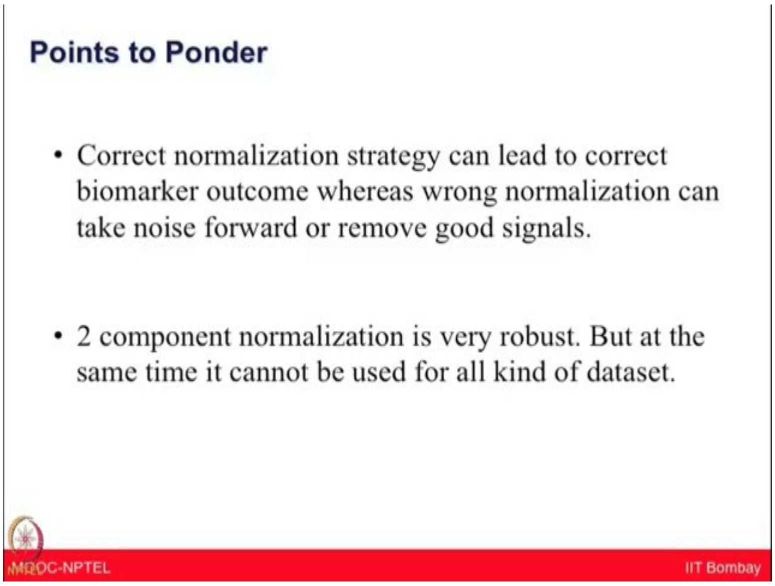


Points to Ponder

- Normalization should be done once for any dataset. The strategies differs with data to data.
- Normalization strategy depends on distribution of data, tail of the distribution and also data size.
- Both Centering and Scaling is done for a dataset it allows robust standardization strategy.

MOOC-NPTEL IIT Bombay

(Refer Slide Time: 30:15)



Points to Ponder

- Correct normalization strategy can lead to correct biomarker outcome whereas wrong normalization can take noise forward or remove good signals.
- 2 component normalization is very robust. But at the same time it cannot be used for all kind of dataset.

MOOC-NPTEL IIT Bombay

So, I hope you have learnt and appreciated how the big data obtained from genomic and proteomic technologies could be very meaningful, but still to obtain the relevant insights; you need to normalize the data.

And today's lecture by Dr. Mani has given you and illuminated you with more thoughts about how to normalize your data, how normalization strategies differ from one to other data set type. You cannot just apply the same normalization for all type of data sets possible, even the data which was obtained from the mass spectrometry or proteomics experiments. You have also heard about the centering and a scaling, when it should be done and how it could actually help you to obtain robust standardization strategies.

I hope you also studied, how the correct normalisation strategy can lead to address the correct biological question or even to correct biomarker identification. And, these outcomes could be very wrong if your baseline was wrong; if your normalisation was not correct to begin with. Therefore, without knowing about these issues and planning to rectify the issues, by using the right way of normalization becomes very crucial.

You also heard that under which context you should do normalization and whether we should apply that you know at with the raw data only once, for the whole data set. You also learnt about the importance of two component normalization and its you know robustness. At the same time you have probably heard the two component normalization cannot be used for all kinds of datasets.

So, again the context in which the normalization strategy should be utilized depends on the distribution of data, the tail of distribution as well as the data size. We will continue the next lecture again by Dr. D. R. Mani and in the next lecture; you will be given concepts and exposure of the importance for the batch correction, as well as the missing values imputation in data analysis.

Thank you.