Chapter 16

Statistics in Experimental Design, Preprocessing, and Analysis of Proteomics Data

Klaus Jung

Abstract

High-throughput experiments in proteomics, such as 2-dimensional gel electrophoresis (2-DE) and mass spectrometry (MS), yield usually high-dimensional data sets of expression values for hundreds or thousands of proteins which are, however, observed on only a relatively small number of biological samples. Statistical methods for the planning and analysis of experiments are important to avoid false conclusions and to receive tenable results. In this chapter, the most frequent experimental designs for proteomics experiments are illustrated. In particular, focus is put on studies for the detection of differentially regulated proteins. Furthermore, issues of sample size planning, statistical analysis of expression levels as well as methods for data preprocessing are covered.

1. Introduction

Sometimes, today's bioanalytical research is accompanied by the phantasm that the more data is recorded within an experiment the bigger will the cognition drawn from this experiment be. This phantasm is stimulated by the new technological possibilities of measuring simultaneously the expression levels of thousands of molecules as well as by the opulent information stored in databases. The good intentions behind high-throughput experiments are, however, opposed by the fact that the probability of wrong conclusions increases with the number of hypothesis stated in the context of an experiment. Patterson (1) consequently named data analysis the "Achilles heel of proteomics." Preconditions for tenable inferences are well-defined study problems, adequate experimental designs and the correct statistical methods for data analysis.

One of the challenges for the analysis of data from high-throughput experiments is their high-dimensionality, i.e., many

Michael Hamacher et al. (eds.), *Data Mining in Proteomics: From Standards to Applications*, Methods in Molecular Biology, vol. 696, DOI 10.1007/978-1-60761-987-1_16,

features are observed on only a small number of biological samples or individuals. Historically, statistical methods for the analysis of high-dimensional data were refined or even newly developed for gene expression data from DNA microarrays. Because studied problems in proteomics are often very similar to those in genomics, many of these statistical methods can easily be employed for protein expression data, too. An essential difference between gene and protein expression data, however, results from the different bioanalytical technologies which are used for measuring expression levels. Therefore, different ways of data preprocessing are necessary.

A particular question of proteomics is the comparison of expression levels between different types of biological samples, for example, between samples of mucosa and tumor tissue. In Subheading 2, experimental designs for such problems as well as issues of sample size planning are detailed. The presented designs are applicable when comparing two or more independent or dependent categories of biological samples. An example for dependent categories are repeated measurements of the same samples at different points in time. Furthermore, models which include more than one experimental factor are illustrated. Subheading 3 presents necessary steps for the preprocessing of expression levels recorded by mass spectrometry (MS) of 2-dimensional gel electrophoresis (2-DE). Preprocessing is necessary for making the recording of different biological samples comparable. In Subheading 4, the statistical concepts of hypothesis testing and of *p*-value adjustment for multiple testing are detailed, as well as the quantification of expression ratios.

2. Designs and Planning of Experiments

A classical laboratory experiment consists of measuring a dependent (or endogenous) variable under the influence of other independent (or exogenous) experimental factors. In proteomics experiments, the dependent variable is usually the expression level of a protein (i.e., a metric variable), whereas the independent variables may be either categorical (e.g., group membership or disease state) or also metric (e.g., age). In the following, we regard different experimental designs, starting with the most simple one, which is given by studying one experimental factor with two categories, and turn then toward several further aspects of experiment planning, such as sample size calculation and randomization. We regard especially experimental designs for 2-D DIGE gels. Further designs for experiments with this particular type of gels also are presented in (2, 3).

2.1. One Experimental Factor with Two Categories

One of the most frequent problems in proteomics is the comparison of expression levels from two distinct types of biological samples, for example, cell lines under two different experimental conditions or tissue samples from diseased and healthy individuals. These experiments have thus only one categorical experimental factor with two categories. In the just mentioned examples, all samples are independent from each other. One can, however, also consider the case of dependent biological samples, for example, tumor tissue and mucosa from the same individual or a cell line sample measured at two different points in time. The goal of such experiments is to find those proteins which are significantly up- or downregulated in the one category of samples compared to the other one. The preprocessing and analysis of the resulting data is described in Subheadings 3 and 4. In the following, the concrete handling of these designs within 2-DE and MS experiments is given.

In a classical 2-DE approach, simply one gel is prepared per sample. When using the so called DIGE approach (4) instead (where two or more samples, labeled by different fluorescent dyes, can be studied on the same gel), the experimentator has to distinguish between experiments with independent and those with dependent samples. In the latter case, i.e., when two samples per experimental unit (individual) are studied, both samples can be prepared onto the same gel, and the ratios of expression levels are used for statistical analysis. When regarding independent samples instead, an internal standard (comprised of a mixture of all samples included in the experiment) is additionally incorporated, and the ratios of expression levels from the true samples to those from the standard are analyzed. In the case of independent samples, two types of experimental settings can be considered when using DIGE gels. In the first setting, each sample is prepared together with the internal standard on one gel. This design is especially recommended when samples sizes are very different for the two categories of the experimental factor. Particularly, when samples sizes are equal, it is also possible to put two samples – each representing one of the two categories of the experimental factor – together with the internal standard onto one gel (three different fluorescent dyes are used, here). This second setting needs less gels than the first one; it is, however, necessary that the two different samples for each gel are assigned together randomly. Procedures for randomization are described below in this section.

MS experiments are performed very similar. In classic approaches, each sample is recorded in one MS run. Newer approaches which incorporate an isotope-labeling can analyze two samples – labeled by masses of different weight – in one run (5, 6). When studying dependent samples using such isotope-labeling approaches, again ratios of observed intensities are taken for analysis. When having samples from two independent groups,

it is again necessary to match two samples – one of each group – randomly for one MS run.

2.2. One Experimental Factor with More than Two Categories

In some experiments which study the influence of one experimental factor, more than two categories are studied. Stühler et al. (7), for example, compared expression levels in brains of mice at different developmental stages, embryonic, juvenile, and adult. When using DIGE gels, it is recommended to put always only one sample together with an internal standard onto one gel. Only if combinations of categories are assigned randomly to a gel, it is also possible to put more than one sample onto the same gel.

2.3. Two or More Experimental Factors

Let us next regard experiments, where two categorical experimental factors are to be studied. It is then necessary to distinguish between designs with a cross-classification and those with hierarchical classification.

In a cross-classified experiment, each category of the one factor is combined with each category of the other factor. Assume, for example, that it is desired to observe the effect of two different treatments A and B on the expression levels in samples from a certain cell line. We have thus two experimental factors, A and B, each with two categories, treated and not treated. One can then prepare the samples under four different conditions: (1) not treated, (2) only treated with A, (3) only treated with B, and (4) treated with A and B.

In a hierarchical design, not each combination of categories from the two factors is studied. Let us consider a study with two cohorts of patients, where each cohort is treated with a different therapy (thus, factor A has two categories: therapy one and two). As second factor B, consider "diabetes mellitus status," with the two categories "present" and "not present." It is obvious, that a patient can neither be studied under each category of factor A nor under each category of factor B, here.

One can consider course experimental designs with even more than two experimental factors, also in cross-classified and hierarchical settings, however, these designs are seldom studied in proteomics.

2.4. Repeated Measures Designs

A special type of experiments is when expression levels are studied multiple times on the same sample, but under different conditions. These designs are called repeated measures designs. Basically, the above detailed design with one experimental factor of two categories is a repeated measures design if the samples from the two categories are dependent, for example, if expression levels are studied in tumor and mucosa of the same patients. A frequently used repeated measures design is usually given if one experimental factor is the time, where the categories of this factor are different points in time. Sitek et al. (8) studied, for example,

cell lines at several hours after the treatment. The dependence structure of such measurements has to be taken into account in the analysis of such experiments.

2.5. Randomization

In all of the above-described experiments, the experimenter is usually only interested in the effects of the intentionally incorporated factors. It can, however, happen that a studied factor is overlapping with another uninteresting one. Assume, for example, that protein expression in the liver of mice from a treatment group is compared with that of an untreated control group. And suddenly, the experimentator (after he has spent a lot of time with collecting and preparing samples) gets aware that all treated mice were male and all untreated individuals were female. Is the experimentator then studying the effect of treatment or of that of gender? (Yes, such disasters happen!)

How can such mistakes be avoided in the planning of an experiment? Particularly, when studying treatment effects, the probability of incorporating undesired overlapping effects can be diminished by assigning the samples to the different categories of the treatment factor randomly (by the way: "randomly" is not the same as "arbitrarily"!). An example is given in the notes section.

2.6. Sample Size Calculations

Because sample and gel preparation is expensive and time consuming, an important question when planning a proteomics experiment is how many samples are needed for a particular experiment. This question can be stated more precisely by the question "How many samples are needed to detect an effect of a certain size?" Consider, for example, a design for comparing two categories of samples and a very small expression change of a particular protein between the two categories is supposed to cause overall strong changes within the studied biological system. A considerable larger number of samples is then necessary to detect this small effect than for detecting a very obvious and big effect. Besides the size of the effect that is to be detected, the variance of the expression levels influences the number of samples, too. The higher the variance, the harder it becomes to detect an effect. Both, the influence of the effect size and that of the variance onto the necessary sample size are exemplified in Fig. 1.

When calculating the appropriate sample size for an experiment, it is therefore necessary a) to specify the size of effect that is desired to be detected and b) to know something about the variance of expression levels. Knowledge about the variance can only be earned from earlier studies or, for example, from a small pilot experiment. With this information, one can calculate the so called power, which is the probability that a truly existing effect is detected by a statistical test (see Subheading 4). Let us regard the example in Fig. 2, where power curves are plotted under the assumption that the variance of the log-transformed expression

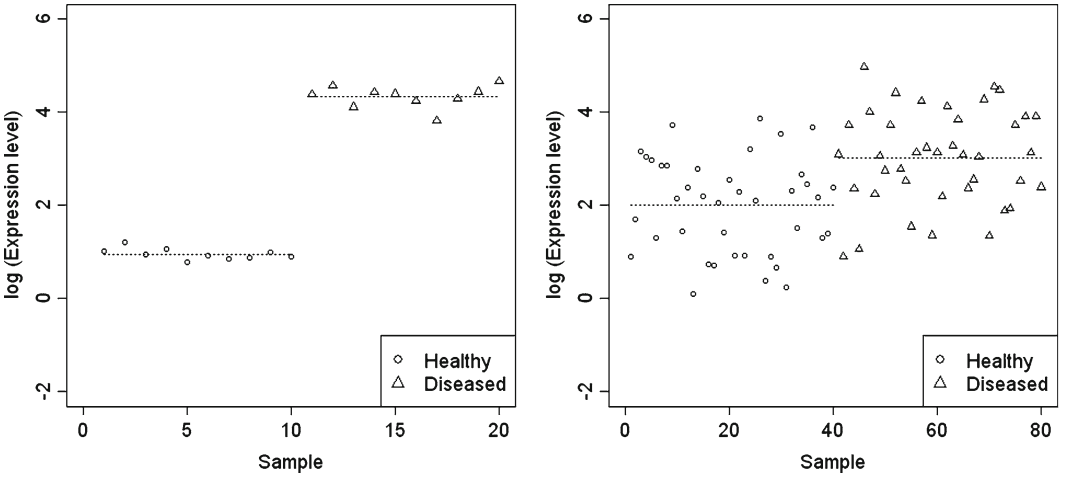

Fig. 1. A small sample size per group is sufficient to detect a big group effect when expression levels scatter very little (*left*), while larger sample sizes are necessary to detect a very small effect or when expression levels scatter very much (*right*).

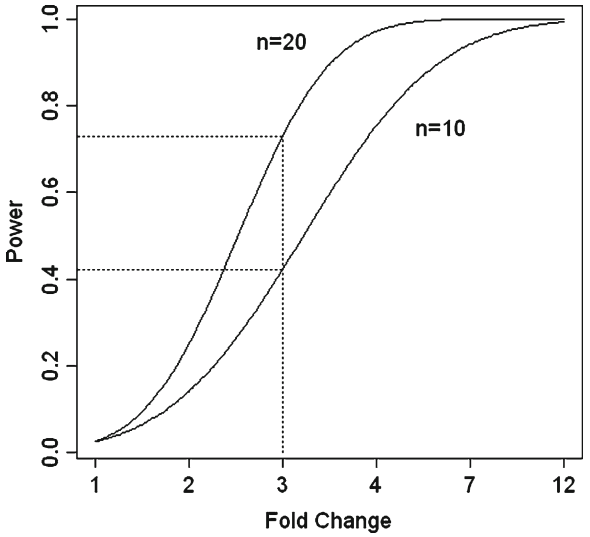

Fig. 2. Theoretical power curves for two different samples sizes n per group. The power is the probability that a particular true fold change is detected by a statistical test.

levels is 1.2 and for two different samples sizes, $n = 10$ and $n = 20$ per group. In this setting, a true 3-fold expression change can be detected with a probability of 0.42 when using 10 samples per group and with a probability of 0.71 when using 20 samples per group. For practical aspects of power calculation, see the notes section.

Another way of determining an appropriate sample size is to control a prespecified false discovery rate (9).

3. Data Preprocessing

Before starting with the concrete statistical analysis, electrophoretic and mass spectrometric recordings must be preprocessed. The raw results of 2-DE experiments are digital images which contain protein spots and the raw result of a mass spectrometric experiment is a mass spectrum with the m/z ratio on the abscissa and the intensity on the ordinate. For the former type of experiment, preprocessing starts with a specified automatic spot detection algorithm and by summarizing the pixel values within a spot boundary as a measure of abundance. For the latter one, a peak detection algorithm is carried out first and then the intensity values within the start and end point of a peak are summarized as a measure of abundance (10, 11). The thus obtained expression levels have to be further transformed by several steps as described in the following.

3.1. Variance Stabilization

In nearly all proteomic experiments, it can be observed that the variance of expression values that have been recorded for a protein depends on the average of these values. In , highly expressed proteins have a higher variance than low expressed proteins. It is therefore a common usage to apply a variance stabilizing transformation to the recorded expression levels. Most common transformation functions are the logarithm or the arsinh function. While the logarithm can, however, produce extreme and negative values for very low expressed proteins, the arsinh is positive and more flat in the lower region.

3.2. Normalization

Normalization is a further necessary transformation of expression levels to make the measurements of several gels or MS runs comparable. Particularly, in experiments where the samples of different categories are prepared by different labels (e.g., different fluorescent dyes or mass tags), normalization can also be used to remove labeling-biases and make the different channels comparable.

The two most frequent used normalization methods for proteomics data are quantile normalization (12) and the vsn normalization (13, 14). Quantile normalization shifts the expression levels of all gels or MS runs to have the same quantiles (see notes section). The vsn method uses affine linear mappings for transforming the expression levels. The latter method directly applies the arsinh function for variance stabilization as described above.

3.3. Standardization

Another method for making several gels comparable is the incorporation of an internal standard. This method can only be applied for techniques in which different labels are used and two or more samples are prepared on the same gel or run within the same MS run. One channel is then usually used for the internal

standard – mostly a mixture of all samples studied within an experiment. Expression values from the true samples are then divided by those of the internal standard. An internal standard is redundant if depended samples are directly compared to each other.

3.4. Missing Values Imputation

Particularly, in 2-DE experiments, resulting data matrixes contain a considerable number of missing values because the number of detected spots or of identified proteins is different from gel to gel (15, 16). Most of the classical statistical methods that were invented in the first half of the twentieth century are, however, designed for complete data matrixes, especially the methods for multivariate data. There are a number of ways missing data can be handled. Perhaps the simplest one is to omit all data rows or columns with missing values. That means, however, a loss of statistical power or a loss of information about certain proteins. Another possibility is to impute missing values and to obtain thus a complete data matrix. Several methods for missing values imputation are possible, e.g., the k nearest neighbor method (see notes section) or principal component regression. More sophisticated methods make imputations repeatedly several times and take the mean of all imputations (17).

4. Statistical Analysis

4.1. Statistical Hypothesis Testing

Let us throw again a glance upon the above-described study problems for comparing samples from two different categories of tissues. The goal of such experiments is the detection of differentially regulated proteins. An easy strategy to find those proteins would be to simply compare the average expression level in both categories of samples, separately for each protein. However, because expression levels are measured on a continuous metric scale, a nonzero difference between the average level in both categories can be expected for almost every protein, even for those which are not differentially regulated. How can the analyst now decide, which of the differences are big enough to call the protein differentially regulated, or how can he distinguish those proteins for which the difference is nonzero just by chance from those for which the difference deviates significantly from zero? This decision can be made by performing a statistical test. For performing a statistical test, first a *null hypothesis* is stated (e.g., "Protein x *is not* deregulated") as well as the complementary *alternative hypothesis* (e.g., "Protein x *is* deregulated"). Based on the measured values and eventually some assumptions about their underlying probability distribution either the null hypothesis is maintained or it is rejected in favor of the alternative hypothesis.

Table 1
Comparison of test decision based on experiment and unknown reality

		Unknown reality	
		Protein *is not* deregulated	**Protein *is* deregulated**
Test decision	Protein *is not* deregulated	True negative decision	False negative decision
	Protein *is* deregulated	False positive decision	True positive decision

Because a test decision is generally based on samples that are taken from a bigger population and because the measured quantity has usually a nonzero variance, the decision may fail to hit the unknown reality. In particular, a false positive or a false negative decision is possible (Table 1). Unfortunately, the probability α for a false negative decision and the probability β for a false negative decision are interdependent, and can thus not be decreased simultaneously. The solution is therefore to predefine a tolerable α (also called level of significance) and to control β by calculating the necessary sample size. A quantity that is usually derived by a test is the so called *p*-value. If this value is smaller than α, the null hypothesis is rejected.

4.2. Comparing Two Groups

In the most frequent problem in proteomics, that is comparing expression levels between two different categories of biological samples, one test is performed for each protein. If one can assume that expression levels are normally distributed, the so called t-test can be used. If expression levels are assumed to be non-normally distributed, e.g., if they show a very skewed distribution, one should rather use the nonparametric Mann-Whitney-U test (MWU). Both, t-test and MWU test offer versions for dependent and independent categories.

4.3. Multiple Hypothesis Testing

When performing thousands of statistical tests simultaneously (i.e., for thousands of proteins), there will usually be a high number of positive test decisions which are made just by chance, though the true situation is not positive. Naturally, these test decisions are false positive ones. How can the number of false positives be diminished? One solution to this problem is to be more conservative when testing. For that purpose, *p*-values can be adjusted in the sense of certain error rates (18), for example, the family-wise error rate (FWER) or the false discovery rate (FDR). The FWER is defined as the probability of having at least one false positive test decision among all test decisions. The FDR,

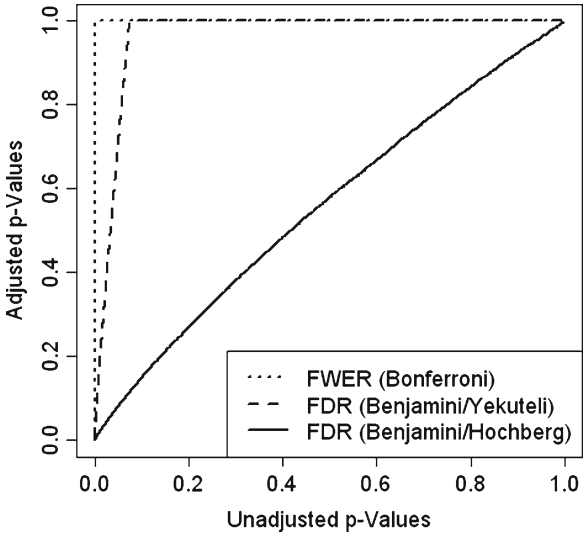

Fig. 3. Relation between unadjusted and adjusted *p*-values.

on the other hand, is the portion of false positives among all positives. An algorithm for adjusting *p*-values with regard to the FWER was given by Bonferroni. For controlling the less strict FDR, there are two different algorithms. One of them assumes that all hypothesis are independent (19) and a more liberal one puts no assumption onto the correlation structure of the hypothesis (20). Adjusted *p*-values from a cell line study of adenocarcinoma (21) are plotted versus the raw *p*-values in Fig. 3. While there seemed to be many significantly regulated proteins when using the unadjusted *p*-values, the FWER- and FDR-adjustments dramatically reduced the set of significant features and thus the number of false positive findings. Formulas for adjusting *p*-values are given in the notes section.

4.4. Analysis of Variance or Covariance

In Subheading 2, we have mentioned experimental designs in which a factor can have more than two levels or in which more than one factor is included. Data from such experiments can statistically be evaluated by analysis of variance (ANOVA) methods or by analysis of covariance (ANCOVA). In both methods, a dependent metric variable (here, these are the expression levels) is related to one or several independent experimental factors. If all independent variables are categorical (e.g., group membership, gender), ANOVA is used. If there is also one or more independent metric variables, ANCOVA is used instead. For each of the independent variables, one statistical test is performed, i.e., one *p*-value is produced. If that *p*-value is smaller than the significance level, the associated variable is supposed to have a significant influence onto the dependent variable. Besides the main effect

given by the independent variables, it is also possible to study interactions between these effects. Assume, for example, that there are two independent categorical factors included in the experiment, each on two levels: group (healthy, diseased) and gender (male, female). A significant interaction between group and gender indicates that the strength of a significant group effect is different within the two levels of gender. In extreme situations, an interaction between two independent variables can mean that the effect of factor A is inverse between the two levels of factor B.

Particular analysis of variance methods for repeated measures designs are, for example, detailed in (22) in the case of a normal distributed dependent variable. Because protein expression levels are often not assumed to be normally distributed, one should also consider nonparametric methods as detailed in (23). The most difficult problem in the analysis of repeated measures designs is to set the correct assumption of the correlation matrix. Different forms for this matrix can be considered. Let us take again the example that protein expression is repeatedly measured at several subsequent points in time. A simple assumption for the correlation structure between the studied points in time is that there is an equal correlation between each of two points in time (this structure is called compound symmetry). More realistic, is however, that points in time that are more distant from each other have a smaller correlation than those that are less distant (autoregressive structure). In some situations, an unstructured correlation matrix is assumed.

4.5. Fold Change and Confidence Intervals

Using statistical tests, it is possible to conclude that a protein is significantly up- or downregulated. One goal of proteomics is furthermore to quantify the strength of regulation, which is usually done by a ratio estimate, for example, the fold change. The fold change is defined as the ratio of the average expression between two categories of an experimental factor. Because expression levels are usually log-transformed, the ratio becomes than a difference. In the context of the fold change, it is important to report this quantity always in combination with a confidence interval (21). A confidence interval covers the true expression change with a probability of $(1-\alpha)$. Using a confidence interval, one can compare the importance of proteins with the same fold change. Assume that protein X and protein Y both have a fold change of 2. For protein X, however, the confidence interval is given by [0.6, 2.8] while the confidence interval for protein Y is given by [1.7, 2.2], i.e., the latter confidence interval is much smaller but with the same level of confidence than the former one. For protein X, it is then not really possible to conclude that it is truly up regulated because the lower bound of the interval is smaller than 1. For protein Y instead, it seems very likely that is up regulated with a high confidence.

5. Notes

5.1. Randomization

In order to avoid the overlapping of effects from the studied experimental factors with other uninteresting effects, a random assignment of experimental units to the study groups is important. Assume, for example, that the effect of one treatment is to be studied on the expression levels in a cell line. Five samples are to be assigned to each group, the treatment and the control group. For random assignment, follow the next steps:

1. Generate a list of ten random numbers (e.g., from a standard normal distribution) and assign ranks 1–10 to these numbers.
2. All samples with rank 1–5 are treated and all with ranks 6–10 are not treated (Table 2).

5.2. Sample Size Calculations

When testing for differential expression between two groups, a *t*-test is usually carried out for each protein. The power of the t-test is the probability that the test detects a certain log(fold change) under a fix sample size and with a given variance of the expression levels. For determining an appropriate sample size, proceed as follows:

1. For each protein in the data from a pilot sample, calculate the variance of its preprocessed expression levels. Power can, for example, be calculated for the minimum, median, or maximum of all variances.
2. Calculate the power for different sample sizes and for different log(fold changes) using the variance estimates from the pilot sample. It is recommended to use a statistical software tool for calculating the power (e.g., the free software R from www.r-project.org).
3. Choose that sample size for your experiment which yield the desired power.

Table 2
Random assignment of treatment or nontreatment to ten samples of a cell line for avoiding undesired overlapping effects

Sample	1	2	3	4	5	6	7	8	9	10
Random Number	1.92	0.15	−0.64	−1.00	−0.83	1.02	0.16	0.54	−0.19	0.34
Rank	10	5	3	1	2	9	6	8	4	7
Treatment	No	Yes	Yes	Yes	Yes	No	No	No	Yes	No

5.3. Quantile Normalization

Assume that your data matrix A consists of m columns (representing gels) and n rows (representing proteins). For making gels comparable, the following steps of quantile normalization can be applied (directly cited from (12)):

1. Sort each column of A, yielding a new matrix A_{sort}.
2. Calculate the means across rows and assign this mean vector to each column of A_{sort}, yielding the matrix M.
3. Rearrange M to have the same ordering as A, yielding the normalized matrix A_{norm}.

This algorithm is also implemented in the "limma"package for the software R (available from www.bioconductor.org).

5.4. Missing Values Imputation

Especially, 2-DE produces data matrixes with many empty entries. To impute these missing values, one can use the k-nearest neighbor method:

1. Calculate the correlation or distance between the expression levels of each pair of proteins by using the available values.
2. Assume that the expression level of protein i in gel j is missing. Determine the k nearest proteins (neighbors) to protein i (according the distance or the correlation). Calculate the mean of the expression levels of these k neighbors in gel j and use it to fill the gap. A k between 10 and 20 has been recommended by Jung et al. (16).

5.5. Adjusting of p-Values

When searching for differentially regulated proteins, p-values should be adjusted to avoid a too high number of false positives. Assume that the raw p-values for n proteins are $p_1, \ldots, p_n$. A very strict adjustment is given by the Bonferroni method to control the FWER:

$$p_i(\text{adjusted}) = \min\{1, n \cdots p_i\} (i = 1, \ldots, n).$$

A less strict method is the FDR-procedure of Benjamini and Hochberg (19).

1. Take the n ordered p-values: $p_{(1)} \le p_{(2)} \le \ldots \le p_{(n)}$.
2. The FDR-adjusted p-values are given by $p_{(i)}(\text{adjusted}) =$

$$\min_{k=i,\ldots,n} \left\{ \min\left(\frac{n}{k} p_{(i)}, 1 \right) \right\}$$

$$(i = 1, \ldots, n).$$

Other adjustment procedures are implemented in the *p.adjust* method of the R-package "stats."

References

1. Patterson SD (2003) Data analysis – the Achilles heel of proteomics. Nat Biotechnol 21:221–222
2. Karp NA, McCormick PS, Russell MR, Lilley KS (2007) Experimental and statistical considerations to avoid false conclusions in proteomic studies using differential in-gel electrophoresis. Mol Cell Proteomics 6:1354–1364
3. Fodor IK, Nelson DO, Alegria-Hartman M, Robbins K, Langlois RG, Turteltaub KW et al (2005) Statistical challenges in analysis of two-dimensional difference gel electrophoresis experiments using DeCyder. Bioinformatics 21:3733–3740
4. Ünlü M, Morgan ME, Minden JS (1997) Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. Electrophoresis 18:2071–2077
5. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol 17:994–999
6. Ross PL, Huang YN, Marchese JN et al (2004) Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Mol Cell Proteomics 3:1154–1169
7. Stühler K, Pfeiffer K, Joppich C, Stephan C, Jung K, Müller M et al (2006) Pilot study of the Human Proteome Organisation Brain Proteome Project: Applying different 2-DE techniques to monitor proteomic changes during murine brain development. Proteomics 6:4899–4913
8. Sitek B, Apostolov O, Stühler K, Pfeiffer K, Meyer HE, Eggert A, Schramm A (2005) Identification of dynamic proteome changes upon ligand activation of trk-receptors using two-dimensional fluorescence difference gel electrophoresis and mass spectrometry. Mol Cell Proteomics 4:291–9
9. Cairns DA, Barrett JH, Billingham LJ, Stanley AJ, Xinarianos G, Field JK et al (2009) Sample size determination in clinical proteomic profiling experiments using mass spectrometry for class comparison. Proteomics 9:74–86
10. Boehm AM, Pütz S, Altenhöfer D, Sickmann A, Falk M (2007) Precise protein quantification based on peptide quantification using iTRAQ™. BMC Bioinform 8:214
11. Jeffries N (2005) Algorithms for alignment of mass spectrometry proteomic data. Bioinformatics 21:3066–3073
12. Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density Oligonucleotide array data based on bias and variance. Bioinformatics 19:185–193
13. Huber W, Heydebreck A, von Sültmann H, Poustka A, Vingron M (2002) Variance stabilization applied to microarray data calibration and the quantification of differential expression. Bioinformatics 18:S96–S104
14. Kreil DP, Karp NA, Lilley KS (2004) DNA microarray normalization methods can remove bias from differential protein expression analysis of 2D difference gel electrophoresis results. Bioinformatics 20:2026–2040
15. Jung K, Gannoun A, Sitek B, Meyer HE, Stühler K, Urfer W (2005) Analysis of dynamic protein expression data. RevStat-Stat J 3:99–111
16. Jung K, Gannoun A, Sitek B, Apostolov O, Schramm A, Meyer HE et al (2006) Statistical evaluation of methods for the analysis of dynamic protein expression data from a tumor study. RevStat-Stat J 4:67–80
17. Schafer JL, Graham JW (2002) Missing data: our view of the state of the art. Psychol Meth 7:147–177
18. Dudoit S, Shaffer JP, Boldrick JC (2003) Multiple hypothesis testing in microarray experiments. Stat Sci 18:71–103
19. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc Ser B 57:289–300
20. Benjamini Y, Yekutelli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat 29:1165–1188
21. Jung K, Poschmann G, Podwojski K, Eisenacher M, Kohl M, Pfeiffer K et al (2009) Adjusted confidence intervals for the expression change of proteins observed in 2-dimensional difference gel electrophoresis. J Proteomics Bioinform 2:78–87
22. Diggle PJ, Liang K-Y, Zeger SL (1994) Analysis of longitudinal data. Clarendon Press, Oxford
23. Brunner E, Domhof S, Langer F (2002) Nonparametric analysis of longitudinal data in factorial experiments. Wiley, New York