2012 IACSIT Hong Kong Conferences
IPCSIT vol. 30 (2012) © (2012) IACSIT Press, Singapore

The Statistics Algorithm For Eliminating the Excrescent Data of Protein Quantification

Chao Ma, Shuai Liu, JunHua Chen and XiaoZhou Chen[+]

School of Mathematics and Computer Science, Yunnan University of Nationalities, Kunming650031, China

Abstract. The accuracy of protein quantification is essential in analysis of biological system. However, it is unenviable that some excrescent data causing the wrong results. So in order to ensure the accuracy of protein quantification, we apply Grubbs test, Ronser test and Med_FD test for analysis the excrescent data in protein quantification. And then we compare several results in which we have analysis with these methods respectively. The results indicate that Grubbs test possesses accuracy effects better than the others.

Keywords: excrescent data; eliminated method; normality test; protein quantification.

1. Introduction

In the post-genomic era, protein quantification analysis has become an interested topic in the field of biological system. It has been shown that the different level of protein expression is closely related to the development of disease, such as cancer [1], viral and bacterial infections [2], and heart disease [3]. So protein quantitative information will help us to explore factors in protein which related to disease.

During the last decade there is a lot of protein quantification techniques have been developed, such as stable isotope labeling of amino acids in cell culture (SILAC), tandem mass tagging (TMT), isobaric tags for relative and absolute quantification (iTRAQ). The development of isotopic labeling technique can reduces the sample's complexity and the relative proteins' quantity can be determined by those corresponding peptides quantity in which comparing the summed peak's intensity. In this paper we used itraq-labeling techniques and this labeling technique is based on a set of four isobaric reagents, each of it comprises three groups: reporter, balance, and reactive groups. The report groups with signature ions at m/z 114, 115, 116, and 117 respectively, the balance groups with m/z 31, 30, 29, and 28 respectively, so there is no increased complexity at the MS level.

In protein quantification, there is a lot of error resulted from operation and equipment inaccuracy [4]. In this study, we find that there existed some abnormal data caused from errors. However, we should note that it is not easily removed from data; it might be results to the wrong results. From this analysis, we applied an efficient statistics approach to eliminate the excrescent data for the accuracy of the results.

2. Materials and Normality Test

This sample's organism is Leptospira interrogans and it was prepared with the following way: leptospiras were exposed to starvation, heat shock 42 deg 1 hour or treated with antibiotics (cipro) for 24 hours or untreated (control). The cells were harvested and the extracted proteins were digested. The peptides were then itraq labeled according to the following 114-control, 115-starvation, 116-heat shock and 117-antiobotic. The samples were then combined and cleaned up before analyzed by LC-MADEL-MS/MS. The sample was injected 5 times mzXML files were created by the 'Michigan software' and database search was done accordingly: Mass accuracy of 1 Da for the precursor mass. For the searched I used fixed Lysine

[+] Corresponding author. Tel.: +86 871 5913120.
E-mail address: *ch_xiaozhou@yahoo.com.cn.*

and Cysteine modification (114.10 @ K, 45.9877 @ C), variable modification on Methionine (15.9949 @ M) and a fixed modification on the N-term (+114.1). This is contributed by Johan Malmstroem.

We proposed to select the unique peptides corresponding to each protein in protein quantification for our investigation. When the number of peptides in one protein is large enough, the data is consistent with the log-normal distribution [5]. Such as 0.63841, 1.07309, 1.9324, 1.1127, 1.3579, 0.8508, 0.7702, 1.0422, 0.9075, 0.7288, 0.7355, 0.6704, 0.4919, 0.3903, 0.7167, 0.2427, and 3.1176 which come from protein “DNA-directed RNA polymerase beta subunit”.

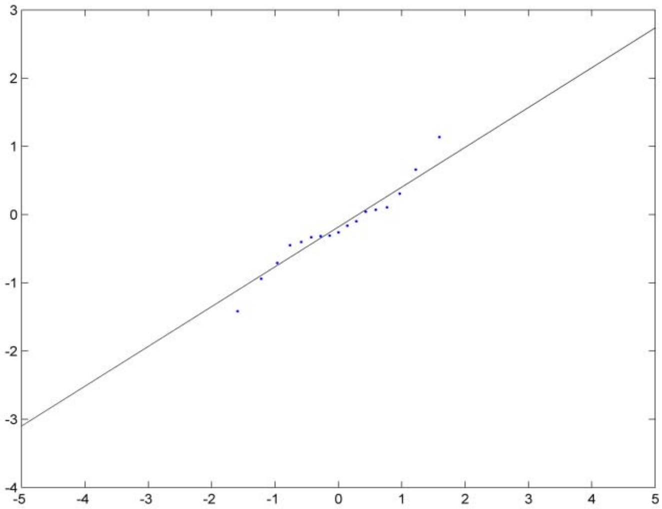

Fig. 1: Data

In Fig.1, the x-axis denotes the inverse function of the normality and the y-axis represents the sorted data. As seen from Fig.1, the log-transformed data is nearly in a straight line. In addition to measuring the data normality with the graphical methods, we also apply Shapiro-Wilk test to measure departure from normality [6], we selected the significance level of α =0.05 in which against the data from Fig.1, and we get the result of statistic W =0.9502 $> W(17, 0.05)$ =0.892. This indicates the log-transformed data are log-normal distributed.

Basing on the premise of the data normally distributed, we analyses the data and verify the existence of excrescent data, and then we proposed three efficient methods for eliminating the excrescent data respectively.

3. Methods

3.1. Grubbs test

Early in 1950 Grubbs test method had been [7] proposed. This approach has a very good effect on tests whether there is excrescent data exists in a set of experimental data.

It is assumed that there are n independent variables $X_1 \le X_2 \le \cdots \le X_n$, from the same normal distribution $N\left(\mu_i, \sigma^2\right)$ respectively, where $i = 1, 2, \cdots, n$. The assumption as follows:

Null hypothesis H_0 : $\mu_1 = \mu_2 = \cdots = \mu_n$

Nonnull hypothesis $\frac{H_1 : \mu_1 < \mu_2 = \cdots = \mu_n}{H_n : \mu_1 = \mu_2 = \cdots < \mu_n}$

Therefore, to test the outlier is to test the H_0 when nonnull hypothesis is H_1 or H_n. For the null hypothesis unsubstantiated, then the minimum or maximum value in the data is treated as outliers. The calculation of statistics as follows:

$$T_1 = \frac{\overline{X} - X_1}{S} \tag{1}$$

If X_1 is excrescent value

$$T_n = \frac{X_n - \overline{X}}{S} \tag{2}$$

If X_n is excrescent value

Where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $S = \left\{ \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \right\}^{\frac{1}{2}}$. The corresponding critical value is:

$$T(\alpha, N) = \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{t^2_{\alpha/N}(N-2)}{t^2_{\alpha/N}(N-2) + N - 2}} \tag{3}$$

Where $t^2_{\alpha/N}(N-2)$ is the function value of t-distribution at significance level of α/N and freedom of $N-2$.

The determination is that: for $X_1 \le X_2 \le \cdots \le X_n$, We selected the significance level $\alpha > 0$ and the number of data N, when $T_1 > T(\alpha, N)$ rejected H_0, and determinate the X_1 is excrescent data, otherwise we can not judge the data is excrescent value. So does the X_n.

3.2. Rosner test

Rosner test [8] is a multiple excluded values testing methods in a single test. For date $X_1 \le X_2 \le \cdots \le X_n$, we firstly set a upper limit value k for suspicious excrescent data, which is the maximum number of excrescent data. The maximum number k can be removed from the data set of sizes n.

It is also assumed that there are n independent variables $X_1 \le X_2 \le \cdots \le X_n$ from the same normal distribution $N\left(\mu_i, \sigma^2\right)$ respectively, where $i = 1, 2, \cdots, n$. The assumption as follows:

Null hypothesis H_0: No excrescent values in the data

Nonnull hypothesis H_l: l excrescent values in the data at least

This method is based on statistics $R_1, R_2, \cdots, R_k$, and defined as:

$$R_1 = (\max |X_i - \overline{X}|)/S \tag{4}$$

Where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$; $S = \left\{ \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \right\}^{\frac{1}{2}}$.

The definition of R_2 like R_1, but the size of data is $n-1$, that is because it's exclude the excrescent data in R_1. We can also define $R_3, \cdots, R_k$ respectively. The critical value of Ronser methods as follows:

$$\lambda_l = \frac{t_{p,k-l-1}(n-l)}{\{[n-l-1+t^2_{p,k-l-1}](n-l+1)\}^{\frac{1}{2}}} \tag{5}$$

Where $p = 1 - \frac{\alpha}{2(n-l+1)}$, $l = 1, 2, \cdots, k$. $t_{p,d}$ represents the function value of t-distribution at significance level of p and freedom of d.

The determination rule is: when $R_i > \lambda_i$, $i = 1, 2, \cdots, k$, the X_i in $R_i = (\max |X_i - \overline{X}|)/S$ is excrescent data, otherwise there is no excrescent data.

3.3. Med_FD test

Med_FD test [9] is a simple and robust method which based on the median value and quartile deviation, and it is not sensitive to the data size. The most feature of this approach is not based on hypothesis testing. The process of this method is that: calculate the median value and quartile deviation for data sets $X_1 \le X_2 \le \cdots \le X_n$, and then set the acceptable region. If there have data beyond the acceptable region, it can be removed as excrescent data. The acceptable region defined as follows ($kl = ku = 2$):

$$X_i \notin [Med - kl * FD, Med + ku * FD] \quad (6)$$

4. Results and Comparison

We selected five proteins and a total of 15 quantitative data sets for this study. We test its normality first (Those data which value size is 0 means do not follow the log-normal distribution). In the condition of normal distribution we applied three methods to test the excrescent values respectively. The results are shown as follows (table 1):

Table. 1: Data

Protein name	Acc.no	Data	Size	Methods		
				Grubbs	Ronser	Med_FD
DNA-directed RNA polymerase beta subunit	YP_000733	115:114	17	no	0.2423	0.2423,3.1126
		116:114	17	no	no	0.5166
		117:114	17	1.5992	1.5992	1.5992
elongation factor EF-2	YP_000262	115:114	12	no	no	no
		116:114	0	---	---	---
		117:114	12	1.4264	1.4264	1.4264
succinyl-CoA synthetase beta subunit	YP_002497	115:114	12	no	0.1254	no
		116:114	0	---	---	---
		117:114	12	no	1.0795	1.0795
Tax_Id=9606 Keratin 10	IPI00383111	115:114	12	no	0.7493	1.8951,0.7493
		116:114	12	no	0.5960	no
		117:114	0	---	---	---
Tax_Id=9606 keratin 1	IPI00220327	115:114	11	no	0.6508	0.6508,1.8692
		116:114	11	no	0.5887	0.5887
		117:114	11	no	no	no

It is not difficult to find that the datasets is generally relatively small in quantitative from table 3, that is because when we select the value of peptide quantitative information that only select the unique peptide corresponding to the protein in order to obtain the accurate results, the non-unique peptides are in the process of protein identification alone. We applied Grubbs methods to detect the small data sets, and only find two excrescent values which are also found in the other methods. Grubbs method is a subset of the other methods. This result indicates that the two values which have been detected must be noise data from errors in protein quantification, so they must be rejected. Compared with Grubbs methods, we found that we remove more values by Ronser test and Med_FD test in small samples. This maybe gets the wrong results for removing too many values. We also could take Grubbs test in big samples and with human experience to ensure the accuracy of results. From above analysis, it indicates that Grubbs test are more suitable for detecting excrescent values in protein quantification.

5. Acknowledgements

This work was partly supported by the Scientific Research Program of Yunnan Provincial Education Department (No. 2011Y215).

6. References

[1] Jones,M.B; Krutzsch,H.; Shu,H.; Zhao,Y.; Liotta,L.A.; Kohn,E.C.; Petricoin,E.F., III. Proteomic analysis and identification of new biomarkers and therapeutic targets for invasive ovarian cancer. *Proteomics* [J], 2002.2(1),PP: 76-84.

[2] Bini, L.; Magi, B.; Marzocchi, B.; Cellesi, C.; Berti, B.; Raggiaschi, R.; Rossolini, A.; Pallini,V. Two-dimensional electrophoretic patterns of acute-phase human serum proteins in the course of bacterial and viral diseases. *Electrophoresis*[J], 1996.17 (3), PP: 612-616.

[3] Scheler, C.; Li, X. P.; Salnikow, J.; Dunn, M. J.; Jungblut, P. R. Comparison of two-dimensional electrophoresis patterns of heat shock protein Hsp27 species in normal and cardiomyopathic hearts. *Electrophoresis*[J], 1999.*20* (18),PP: 3623-3628.

[4] Lin,W.T., Hung,W.N., Yian,Y.H., Wu,K.P., Han,C.L., et al. Multi-Q: a fully automated tool for multiplexed protein quantitation. *Journal of proteome research*[J], 2006. 5, PP: 2328-2338.

[5] Jung K,Gannoun A,Sitek B,Meyer HE,Stuhler K,Urfer W:Analysis of Dynamic Protein Expression Data. *REVSTAT Statistical Journal*[J], 2005.3, PP: 99-111.

[6] Royston, P., Remark AS R94: A Remark on Algorithm AS 181: The W-test for Normality. *Applied Statistics[*J],1995.44, PP: 547-551.

[7] Verma, S. P. and Quiroz-Ruiz, A., Critical values for 22 discordancy test variants for outliers in normal samples up to sizes 100, and applications in science and engineering. *Revista Mexicana de Ciencias Geologicas*[J],2006.23, PP: 302-319.

[8] Rosner, B., Percentage Points for a Generalized ESD Many-Outlier Procedure. *Technometrics*[J],1983.25, PP: 165-172.

[9] Lin HongHua. The robustness data processing method about eliminating outlier. *Journal of China JiLiang University*[J],2004.15(1), PP: 20-24.