Clustering likelihood curves: Finding deviations from single clusters

Claudia Hundertmark[1], and Frank Klawonn[2]

[1] Department of Cell Biology, Helmholtz Centre for Infection Research, Inhoffenstr. 7, D-38124 Braunschweing, Germany
[2] Department of Computer Science, University of Applied Sciences Braunschweig/Wolfenbuettel, Salzdahlumer Str. 46/48, D-38302 Wolfenbuettel, Germany

Abstract. For systematic analyses of quantitative mass spectrometry data a method was developed in order to reveal peptides within a protein, that show differences in comparison with the remaining peptides of the protein concerning their regulatory characteristics. Regulatory information is calculated and visualised by a probabilistic approach resulting in likelihood curves. On the other hand the algorithm for the detection of one or more clusters is based on fuzzy clustering, so that our hybrid approach combines probabilistic concepts as well as principles from soft computing. The test is able to decide whether peptides belonging to the same protein, cluster into one or more group. In this way obtained information is very valuable for the detection of single peptides or peptide groups which can be regarded as regulatory outliers.

Keywords: Clustering, iTRAQ™, LC-MS/MS, likelihood curve

1 Introduction

Comparative analyses between normal and pathological states of biological systems are an important basis for biological and medical research. Therefore, quantitative analyses of biological components, e.g. proteins, are of particular importance. Proteins are the basic components of cells and responsible for most of the processes in organisms. The basic structure of proteins are amino acid chains, which can be digested into smaller chains termed peptides. The totality of proteins, called the proteome, is highly dynamic and can vary significantly concerning its qualitative and quantitative composition due to changed conditions. A common technology for the analysis of the proteome is called liquid-chromatography mass spectrometry (LC-MS/MS).

In the meantime besides protein identification, mass spectrometry enables relative peptide quantitation by LC-MS/MS. One of the most popular technologies for this purpose is called iTRAQ™, which is based on chemical labelling of peptides. iTRAQ™ allows comparative analyses of up to eight proteinogenic samples in parallel (see [1], [2]). After iTRAQ™-labelling of peptides from different samples and subsequently LC-MS/MS analysis a so-called mass spectrum is available for every detected peptide. The obtained mass spectrum contains information on the amino acid sequence as well as information on the relative concentrations of the actually measured peptide in all analysed samples. Concentration of a peptide is given by a so-called "intensity". In order to compare the concentrations of a peptide under different conditions, a ratio is calculated by dividing the intensities measured from the different samples (or conditions). The resulting ratio is termed "expression ratio" and "regulation factor", respectively.

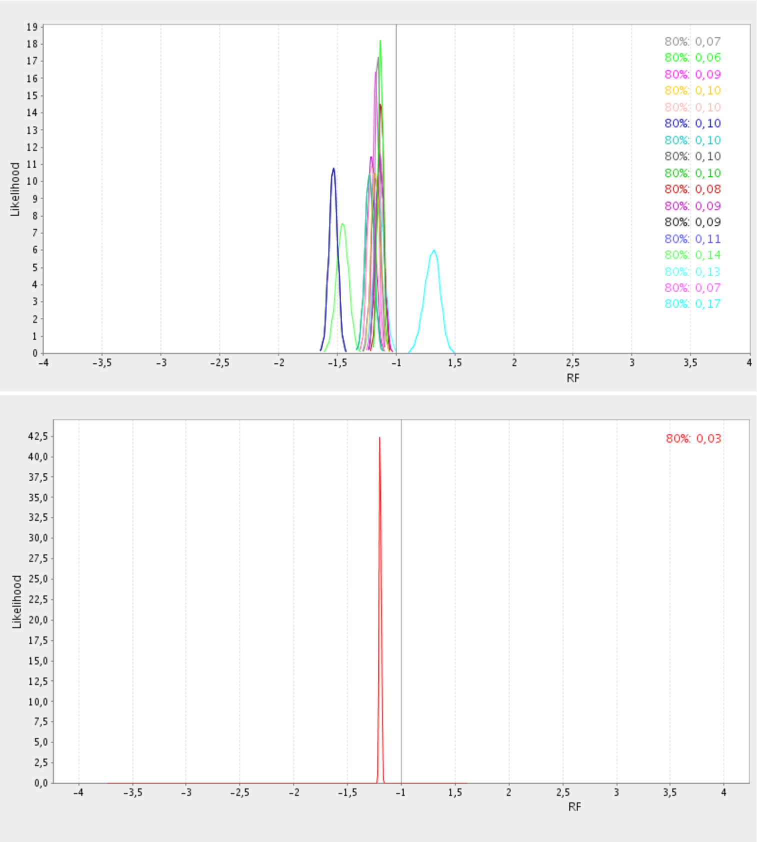

Fig. 1. Likelihood curves represent the most suitable regulation factor of peptides (top) and proteins (bottom), respectively as well as robustness of regulatory information. High iTRAQ™ intensities correlate positively with the robustness and thus result in narrow Likelihood curves on the one hand and in small IR values on the other hand.

In several studies noise was observed in iTRAQ™ analysed data. Particularly if a peptide is detected with low intensities, impreciseness of the calculated ratio is significantly higher than in the case of peptides, that were measured with high intensities. This effect is called "intensity-dependent noise". Based on this observation we developed a mathematical model for the estimation of noise inherent in quantitative data analysed from iTRAQ™-labelled peptides analysed by LC-MS/MS (for details see [3]). Subsequently, we were able to derive several applications from our noise model, e.g. likelihood curves for the calculation of the most suitable regulation factor for single peptides and total proteins as a collective of all contributing peptides ([4]). Furthermore, likelihood curves provide evaluation of robustness of the calculated regulation factor, which again is strongly depending on measured intensities.

Depending on the aim of performed analysis various views on a protein may be useful. Peptides can be visualised separately by individual likelihood curves or they can be combined and visualised by a shared protein likelihood curve. Fig. 1 shows both peptide view (top) and protein view (bottom) for the same protein. Likelihood curves give the likelihoods for regulation factors, which are deviating from the calculated ratio to a greater or lesser extent. X-axis refers to regulation factors, y-axis refers to the likelihood.

Robustness of the underlying data is proportional with both the height and the slope of the produced curve. Furthermore, robustness is represented numerically by the interval of robustness (IR), which is given on the upper right hand side within each plot. IR gives the length of the minimal interval that is covered by 80% of the area beyond each area-normalised likelihood curve (normalised to $\int = 1$). All peptide curves are fairly robust (IR between 0.06 and 0.17) and most of them are fluctuating near -1.2 (no regulation). Robustness of the resulting protein curve of all peptides is extremely high (IR = 0.03) and the regulation factor approximates the regulation factor of the majority of peptide curves.

2 Distance measure

Studying regulatory information of peptides, which is a common practice in biology (especially in proteomics), requires the detection of peptides, that differ in their regulatory characteristics from the remaining peptides of the same protein. Studies like this are important in order to reveal mismatched (by software) peptides or for the investigation of special peptide modifications. For the analysis of large protein samples resulting in very large datasets (several hundreds of proteins) cluster analysis for the automatic identification of proteins consisting of more than one group of peptides concerning their regulatory behaviour is a very helpful means. In the majority of cases all peptides, that belong to the same protein, fluctuate near the same regulation factor and therefore build one cluster. Hence, distinguishing between proteins containing one cluster of peptides and those, which have more than one peptide cluster is the main focus in this approach.

For the following cluster analysis a distance measure d_{ij} giving the distance of two elements i (prototype) and j (peptide j) is to be defined. In order to compare area-normalised likelihood curves i and j the size of the overlapping area a is the distance measure if there is an overlap of curves. Then the distance d_{ij} is given by

$$d_{ij} = \begin{cases} 0 \leq 1 - a \leq 1 & \text{partial overlap of curves } i \text{ and } j \\ 0 & \text{total overlap of curves } i \text{ and } j \end{cases} \tag{1}$$

with a = size of overlapping areas of curves i and j.

In the case of non-overlapping curves i,j d_{ij} is defined by the distance with regard to the scaling of the x-axis which is given by the distance of the highest calculated regulation factor x_{max} of the lower regulated element el_1 and the lowest calculated regulation factor x_{min} of the higher regulated element el_2. Since $d_{ij} = 1$ for curves without overlap, this is the minimum value for non-overlapping curves and must be added to the calculated distance. Therefore, in the case of non-overlapping curves i and j the distance measure d_{ij} is given by

$$d_{ij} = x_{\min}^{(el_2)} - x_{\max}^{(el_1)} + 1 \tag{2}$$

3 Testing for the existence of a single cluster

In search of proteins with significant differences in regulation of related peptides proteins are out of interest for further analyses, which consist of only one cluster. Those proteins are to be identified and can be discarded for following investigation. Therefore, we developed a method for the detection of consistently regulated proteins in order to obtain the remaining proteins consisting of differentially regulated peptides. For these purposes we created a hybrid approach. We combined the probabilistic proceeding for the calculation of protein and peptide likelihood curves with a prototype based fuzzy clustering method .

Our approach for the detection of one-cluster-proteins is based on the generation of a prototype for all regarded peptide likelihood curves and subsequently removing the most distant peptide curve in terms of the prototype curve. In this way the set of likelihood curves is reduced little by little and the distances of the removed curves are plotted by the means of a scatterplot. Analysis of entries within the scatterplot returns the result whether there is only one cluster or not. In detail the procedure is as follows:

First of all, a prototype curve representing all likelihood curves of the protein is calculated. Since the distance measure d_{ij} is based on area overlaps the prototype curve must consist of those parts of all peptide likelihood curves, where most of the peptide curves are overlapping. According to likelihood curves the prototype curve has to consist of connected areas and the likelihood must always be greater than zero at every discretised regulation factor. In detail, calculation of the prototype is done by the detection of the number of overlapping curves for every area that is located below one curve at least. Afterwards, areas, that are part of the highest number of likelihood curves are added to the prototype curve subsequently. This process is finished as soon as the total area of the prototype curve reaches 1.

The second step consists in repeatedly calculating the distances d_{ij} of every peptide likelihood curve and the prototype by application of (1) and (2) and removing the most distant object until there are no likelihood curves left. The respective distances of the removed curves are plotted against the number of iterations within a scatterplot.

Finally, the scatterplot shows whether the protein consists of only one or more different regulatory peptide clusters. In the case of one cluster the entries are arranged homogeneously and close to each other. In the case of multiple clusters on the other hand, there are groups of entries. After removing the last curve of a cluster a significant step is observable in the scatterplot. For automatic detection of steps a threshold is defined: A step is regarded to be significant if the absolute value between the new entry e_{t+1} and the previous entry e_t is greater than the fivefold mean difference $\overline{\Delta_c}$ from the mean value $\overline{x_c}$ within the actual regarded (or lost) cluster. Hence,

$$step = \begin{cases} \text{significant} & \text{if } \left| \mathrm{e}_{\mathrm{t+1}} - e_t \right| > 5\overline{\Delta_c} \\ \text{not significant} & \text{else} \end{cases} \tag{3}$$

It should be noted that the distance of two curves within a cluster is always smaller than 1.

The following figures illustrate this effect: Fig. 2 shows six peptide Likelihood curves for a protein, which are clearly clustering into two groups. The corresponding scatterplot (Fig. 2 insert) shows two groups as well. According to the likelihood plot, every group is composed of three entries. The step between the third and the fourth entry identifies the loss of the last curve from one cluster and is equivalent to the distance of both clusters.

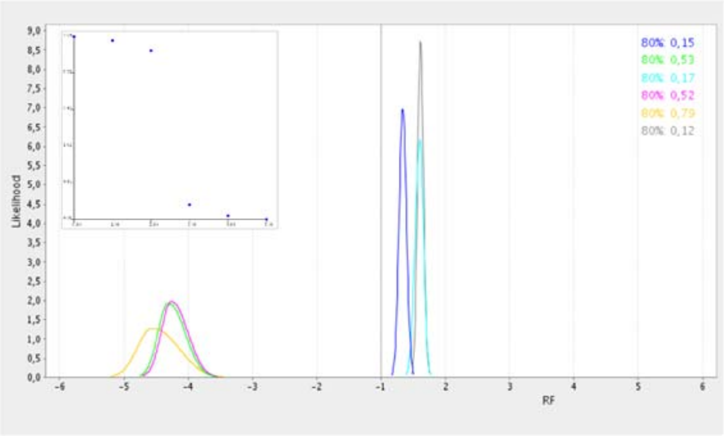

Fig. 2. Likelihood plot consisting of six peptide curves, which are clustering into two groups. Insert: The corresponding scatterplot clearly indicates that the protein is composed of more than one peptide cluster.

Fig. 3 gives an example for a protein, whose peptides are regulated very similarly. Therefore, our method returns that only one cluster is found, which is indicated by one group of entries that are located close to each other. For comparability the scaling of y-axis is the same as in Fig. 2.

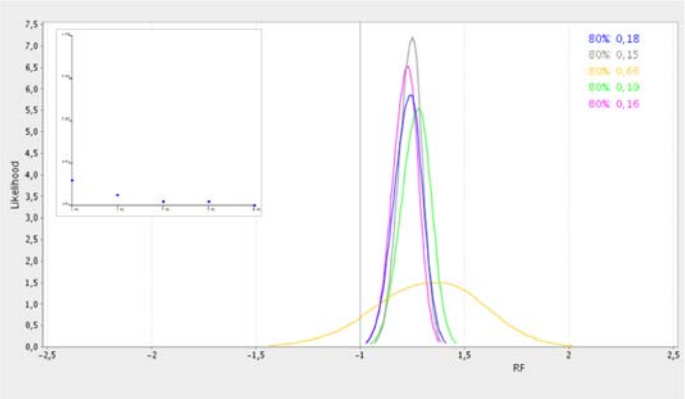

Fig. 3. Likelihood plot consisting of five peptide curves, which are clustering into one group. Insert: The corresponding scatterplot clearly indicates that the protein is composed of one peptide cluster.

Unfortunately, the test is not able to differentiate between different clusters, whose distances to the main cluster are similar. Since the test is using distances without consideration of the localisation of the likelihood curves (lower or higher regulated than the main cluster) two similar distant clusters result in a common step in the scatterplot. Therefore, the test is not suitable for the determination of the number of clusters, but only for the discrimination of one cluster on the one hand and multiple clusters on the other hand. An example is given in Fig. 4: Next to the main cluster (in the middle) are additional clusters on the left as well as on the right. Both additional clusters are located very closely to the main cluster but they are not overlapping. Regarding the inserted scatterplot, it can be observed clearly that four curves are removed in the beginning and after a step the remaining curves are considered. Curves removed first originate from both additional clusters (the left cluster consists of three curves which can not be kept apart visually in the likelihood plot).

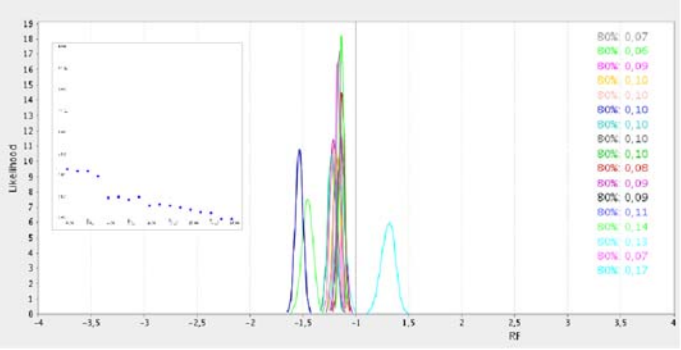

Fig. 4. Likelihood plot consisting of 17 peptide curves, which are clustering into three groups. The leftmost cluster is composed of three curves, the cluster in the middle consists of 13 peptide curves and the rightmost cluster contains one curve. Distances between the main cluster in the middle and the leftmost and the rightmost cluster respectively are more or less equal. Insert: The corresponding scatterplot clearly indicates that the protein is composed of more than one peptide cluster.

4 Conclusion

We have presented an approach for clustering complex data types in the form of likelihood curves. Our method has the advantage that it can cope with small datasets as well as with a small number of cluster, even one cluster. Determining the number of clusters based on validity measures as it is for instance done in fuzzy cluster analysis (see for instance [5,6]) is not suitable for such datasets. Although our approach as it is presented here is only designed to distinguish between one cluster and more than one cluster, our method can also be applied to determine the number of clusters in the following. In case more than one cluster is detected in the dataset, the main cluster corresponding to the likelihood curves after the last significant drop in the scatter plot like in Fig. 2 is removed from the dataset and our method is applied again to the remaining data. A similar idea of subtractive clustering based on removing distant data points, however for standard datasets, but not for likelihood curves, has also been proposed in [7,8].

Acknowledgments. We are very grateful for the support provided by Dr. Lothar Jänsch and Dr. Tobias Reinl.

References

1. Ross, P. L., Huang, Y., N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobsen, A., Pappin, D. J.: Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Mol Cell Proteomics **3** 1154--1169 (2004)
2. Pierce, A., Unwin, R. D., Evans, C. A., Griffiths, S., Carney, L., Zhang, L., Jaworska, E., Lee, C.-F., Blinco, D., Okoniewski, M. J., Miller, C. J., Bitton, D. A., Spooncer, E., Whetton, A. D.: Eight-channel iTRAQ enables comparison of the activity of 6 leukaemogenic tyrosine kinases. Mol Cell Proteomics (2007)
3. Klawonn, F., Hundertmark, C., Jänsch, L.: A Maximum Likelihood Approach to Noise Estimation for Intensity Measurements in Biology. Proc. Sixth IEEE International Conference on Data Mining Workshops ICDM Workshops 2006 180—184 (2006)
4. Hundertmark, C., Fischer, R., Reinl, T., May, S., Klawonn, F., Jänsch, L.: MS-specific noise model reveals the potential of iTRAQ™ in quantitative proteomics. *in preparation*
5. Bezdek, J.C., Keller, J., Krishnapuram, R., Pal, N.R.: Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Kluwer, Boston (1999)
6. Höppner, F., Klawonn, F., Kruse, R., Runkler, T.: Fuzzy Cluster Analysis. Wiley, Chichester (1999)
7. Georgieva, O., Klawonn, F.: Cluster Analysis via the Dynamic Data Assigning Assessment Algorithm. Information Technologies and Control 14—21 (2006)
8. Klawonn, F., Georgieva, O.: Identifying Single Clusters in Large Data Sets. In: Wang, J. (ed.): Encyclopedia of Data Warehousing and Mining. Idea Group, Hershey, 180 – 183 (2006).