Multivariate Time Series Analysis of Metabolite Data from Chinese Hamster Ovarian cell Fermentation

Florian Lipsmeier, Stephanie Esslinger and Oliver Popp

Therapeutic Modalities Informatics – Biostatistics, Pharma Research and Early Development (pRED), Roche Diagnostics GmbH, Penzberg, Germany

Biologics Research, Pharma Research and Early Development (pRED), Roche Diagnostics GmbH, Penzberg, Germany

Abstract

Chinese ovarian hamster (CHO) cells are one of the "workhorses" in modern biotechnology for large-scale production of therapeutic proteins, such as antibodies. Populations of CHO cells grow in a well-defined medium and under well-defined conditions in a fermenter. During fermentation of CHO cells, both a small set of online data are constantly recorded and a discrete set of samples are taken equally distributed over the whole process. These samples are mainly analyzed for their metabolite content and represent the so-called offline measurements. Despite the almost constant conditions during fermentation, there can be a quite high variability in cell growth, process stability and protein production. This may be due to the different proteins produced, different pre-culture conditions before fermentation, mutations or other potentially unidentified factors. We are developing an analysis platform in JMP in order to address different questions around these topics by the analysis of our multivariate time series data. Major tools we use in this context predominantly employ the multivariate analysis methods JMP offers, such as Principal Components Analysis and clustering, and also modeling methods, such as effect screening, partitioning or neural networks. These methods are complemented, for example, by the implementation of different time series similarity measures from the current literature with JMP Scripting Language and the use of existing R packages via the JMP and R connection.

Introduction

Therapeutic proteins such as antibodies become more and more important in drug development. A decade ago only one out of the top ten drugs was a therapeutic protein, whereas it is predicted that until 2014 seven out of the top ten drugs will be therapeutic proteins [1]. The vast majority of therapeutic proteins is made in Chinese ovarian hamster (CHO) cells [2]. This makes them the "workhorses" in modern biotechnology.

Without going into the details of the whole process of therapeutic protein production, it can be summarized by two different sub-processes. In a first step the DNA coding for the therapeutic protein has to be inserted into the genome of a CHO cell. In a second step this CHO cell is then cultivated in a well-defined medium and under well-define conditions in a fermenter, where it replicates and makes the therapeutic protein. Because of its importance in biotechnology, this whole process has been subject of intense research which lead to tremendous improvements in terms of productivity [3]. Despite all these efforts there is still much to learn. For example, we still do not fully know all the important mechanisms inside a CHO cell, which are important for cell growth and therapeutic protein production [4].

With the JMP module described here we focus on the second step of the process, the cultivation of CHO cells in a fermenter. During fermentation several kinds of measurements are collected regularly. There are the so-called online measurements, which comprise all continuous measurements of the fermenter, such as pH, O2 or CO2. The more informative measurements are the offline measurements. On a daily basis samples are taken from the fermenter and these are then in the main part analyzed for their extracellular metabolite content, such as the amino acids, glucose or lactate. Depending on the questions you want to answer you can also go much further and analyze cells in these samples for gene expression or intra-cellular metabolites.

Combining all measurements we get a multivariate time series data set which can hopefully help to explain the outcome of a fermentation and thereby help to improve further fermentation runs. It is important to note that we deal with small time-series of 10-14 measurements showing almost no periodicity. Therefore, the standard approaches of multivariate time series analysis, for example ARMA models, are hardly applicable [5]. As a consequence the JMP module for analyzing CHO cell fermentation data mainly makes use of standard methods from multivariate data analysis such as PCA or clustering. These methods were modified if appropriate following ideas from the research in gene expression or metabolite analysis.

Using JMP as our analysis platform offers different advantages. For one it already comes with many different analysis methods. Furthermore it offers different ways to expand the functionality either by using the scripting language directly to implement new methods or by accessing R and make use of publicly available packages. Finally, our ultimate goal is to provide a JMP module which is not only a collection of analysis methods but can be used by statisticians and also by biologists. This is simplified by the JMP Application builder and the easy way to build up analysis workflows including detailed explanations for every step and resulting in a final report describing all analysis steps and the results.

Material and Methods

The example data set

In order to show the different functionalities of the JMP module, we use real fermentation data which was originally collected in order to assess the variability in a small-scale fermentation process. In this experiment four independent fermentations were done. Each fermentation took about 14 days and at 12 distinct time points samples were taken and analyzed for amino acids, vitamins, glucose, lactate, and several other factors. Though strictly speaking some of these factors are no metabolites, in the following we always speak of metabolites for sake of simplicity. For confidentiality reasons all metabolite names are replaced by letters.

Methods

In this paragraph we briefly describe the methods which are different or modified from the methods offered by JMP.

PCA

Principal component analysis (PCA) is one of the most valuable tools when analyzing multivariate data. Through its dimension reduction capability it enables a first informative view on the data as a whole. We will show in the result section that the multivariate data as produced during fermentation is very well suited for PCA analysis which is due to the strong dependencies between the different factors. After all, they are manipulated by the same process, the metabolism of a CHO cell.

In addition to standard PCA as supplied by JMP, we introduce the dynamic PCA (DPCA) method as described in [8]. We deal with time-series, therefore we cannot assume independency between different samples of the same fermentation run. To accommodate for this, we transform the data set such that every sample consists not only of the measurements at one time point but also of the *d* previous ones. Furthermore we calculate the Hotelling's T^2 statistic for every sample. Following the idea from [8] we use the time-versus-T^2 data in order to find phase shifts during a given fermentation run. A very significant phase shift occurs for example when the CHO cells shift from stationary to exponential growth. Identifying phase shifts allows us to either align data sets from different fermentation runs even if they do not grow synchronously or to analyze each phase separately. Another way to achieve phase separation and alignment is dynamic time warping [11], but we can show that for our kind of data sets DPCA works more stable and accurate. Furthermore, we might detect minor phase shifts beside the main ones which can serve as early indicators to interesting and often not anticipated changes during fermentation.

MEBA

Similar to a web-based tool for metabolomics analysis called MetATT [9] we offer multivariate Bayesian time-series analysis (MEBA) as proposed in [6]. Originally this approach was developed for analysis of microarray time-course data and is available in the R package timecourse from Bioconductor. We introduced slight modifications to the method in order to accommodate for the differences of the data and use the R interface of JMP to access this method. When you can group your data either according to different pre-specified experimental conditions for different fermentation runs or you are for example able to cluster different fermentation runs with respect to the final outcome, you can use MEBA to compare the profiles of the different time series of the metabolites. The outcome is a ranked list of all metabolites which show differences in their temporal profile between the groups of fermentation runs.

ASCA

ANOVA simultaneous component analysis (ASCA) was developed in order to circumvent the problem of traditional MANOVA, which cannot deal with data that consists of more metabolites than samples [7,9,10]. As the name already implies it combines PCA and ANOVA and tries to isolate the variation introduced by different experimental factors. This might reveal hidden relations between these factors and different metabolites.

Results

In this section we highlight some of the possibilities of the JMP module for multivariate time series analysis. The module and JMP itself offer a variety of analysis possibilities and combinations of these.

Here we follow one possible analysis workflow for the example data set to highlight the abilities and advantages of the module.

Step 1: Analysis preparations

After starting the JMP module, you begin by selecting the data set you want to analyze and get two options to go on. You can either start the analysis on the original measurements or you can convert them into consumption rates per cell (see Figure 1).

In either case the first process which is started is an interpolation over the time-series with subsequent prediction of artificial measurements at equidistant time points. This is done by using the spline fitting functionality of JMP via the scripting language. Figure 2 shows a plot of the viable cell densities of the four different fermentation runs after interpolation via spline fitting.

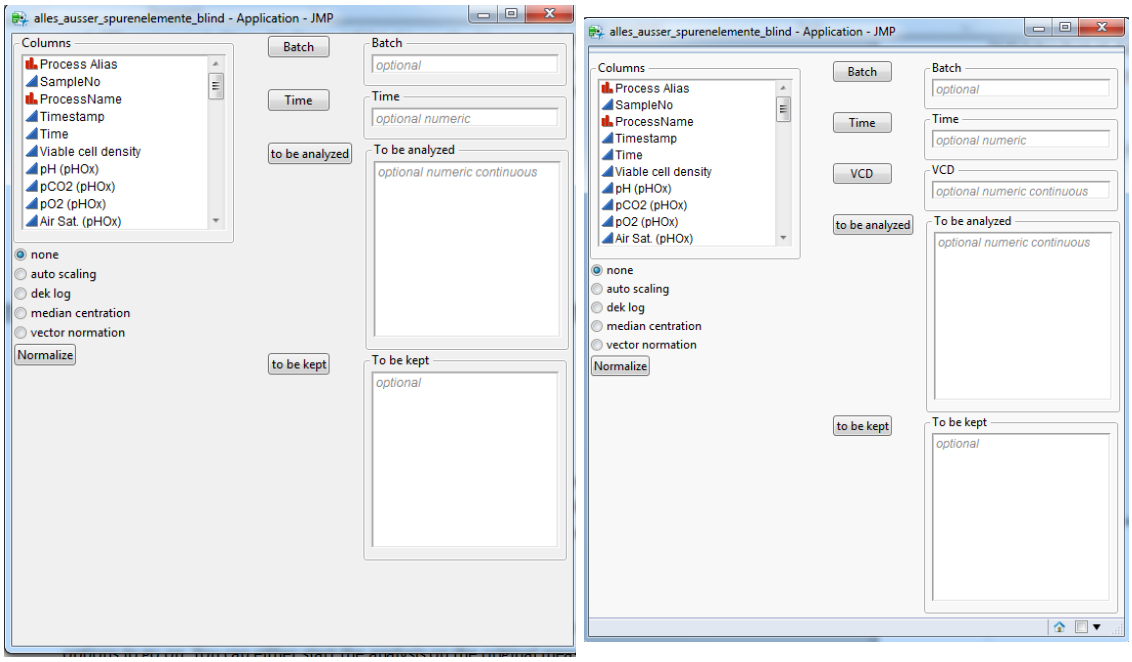

Figure 1 The left figure shows the dialog for starting the interpolation on the original measurements. The right figure shows the dialog for starting interpolation and subsequent consumption rate calculations. The dialog offers also additional possibilities to scale the data.

The user has the possibility to choose the number of points in time that is used for the analysis and thereby defines the granularity of the analysis with respect to changes in time. Besides this obvious advantage, the interpolation is also a necessity to enable a valid comparison of different fermentation runs, especially if they were done independently. As sample collection is no automated process, there are always differences in the time points between fermentation runs. By interpolation we furthermore compensate for missing due to a measurement problem for a certain metabolite. In the case of consumption rate calculation this approach offers the additional advantage that we can approximate a first derivative of a continuous function instead of using differences between measurements of sample points at two consecutive days. Let $C(t)$ be the fitted spline function for a given discrete time series of metabolite concentrations and let $N(t)$ be the fitted spline function for the discrete time series of the viable cell density then the consumption rate can be easily calculated as

$$qP_{metabolite}(t) = C\dot{(t)}/N(t)$$

Drastic changes between two time points either regarding a metabolite concentration or the viable cell density as well as possible measurement errors have not such a strong effect anymore because they are equally distributed over several artificial sample points in a given time interval. Additionally, the spline fitting in itself accommodates for obvious outliers. This leads to smooth instead of sawtooth like lines. See Figure 3 for a comparison between the direct consumption rate calculation versus our proposed calculation at the example of the amino acid asparagine.

Both, analysis of original concentrations and analysis of consumption rates, can be found in the current literature. There are also examples where both kind of data are analyzed together, for example with partial-least square estimation [9]. In the following we concentrate on different possibilities to analyze the original concentrations following the main line of research in the literature. The analysis part for consumption rates is still under development.

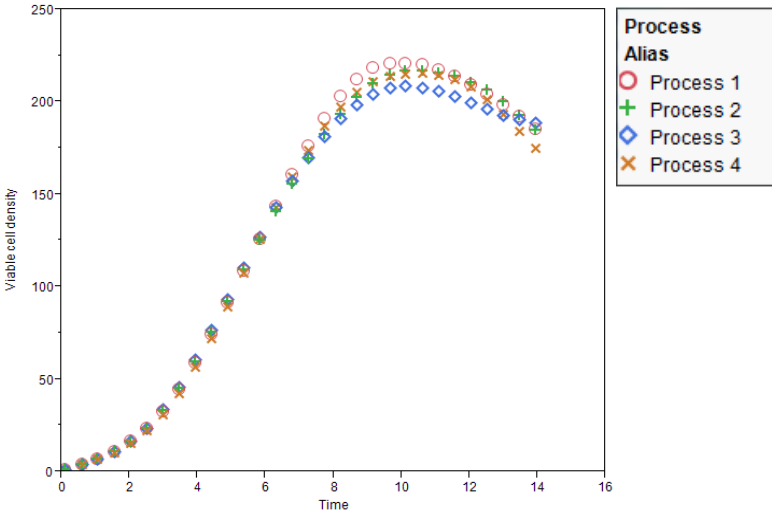

Figure 2 Plot of viable cell densities of the four independent fermentation runs after interpolation via spline fitting.

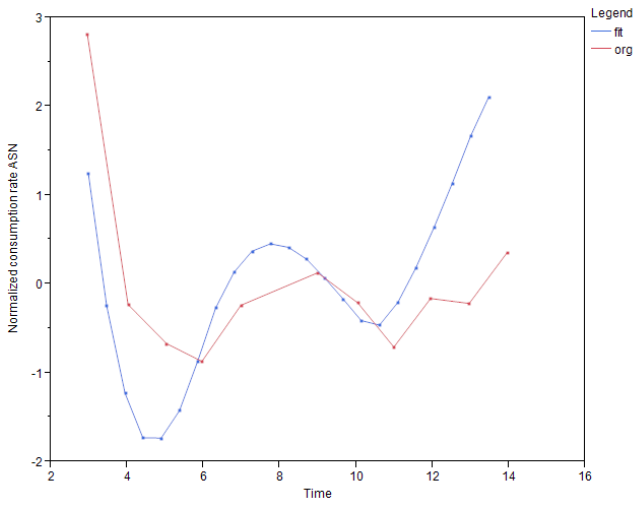

Figure 3 Direct consumption rate calculation versus consumption rate calculation via spline fitting at the example of the amino acid asparagine (ASN)

Step 2: Phase shift identification

Following the spline fitting, the next step in workflow is the identification of phase shifts. The user gets the possibility to choose which measurements should be included and the data is then send to R to perform the DPCA method. As a result a DPCA table with a new column holding the Hotelling's T^2 values is returned. This is complemented by visualizations of the PCA results and time versus Hotelling's T^2 values (Figure 5). Additionally, the method uses the Hotelling's T^2 values in order to automatically detect phase shifts, split the data table accordingly and add a new data table per phase shift. In a new dialog the user gets the option to do a manual phase detection and table splitting if the automatic phase detection did not work satisfactory. Afterwards he can choose to go on with the analysis on the complete data set or only with the data belonging to one of the phases.

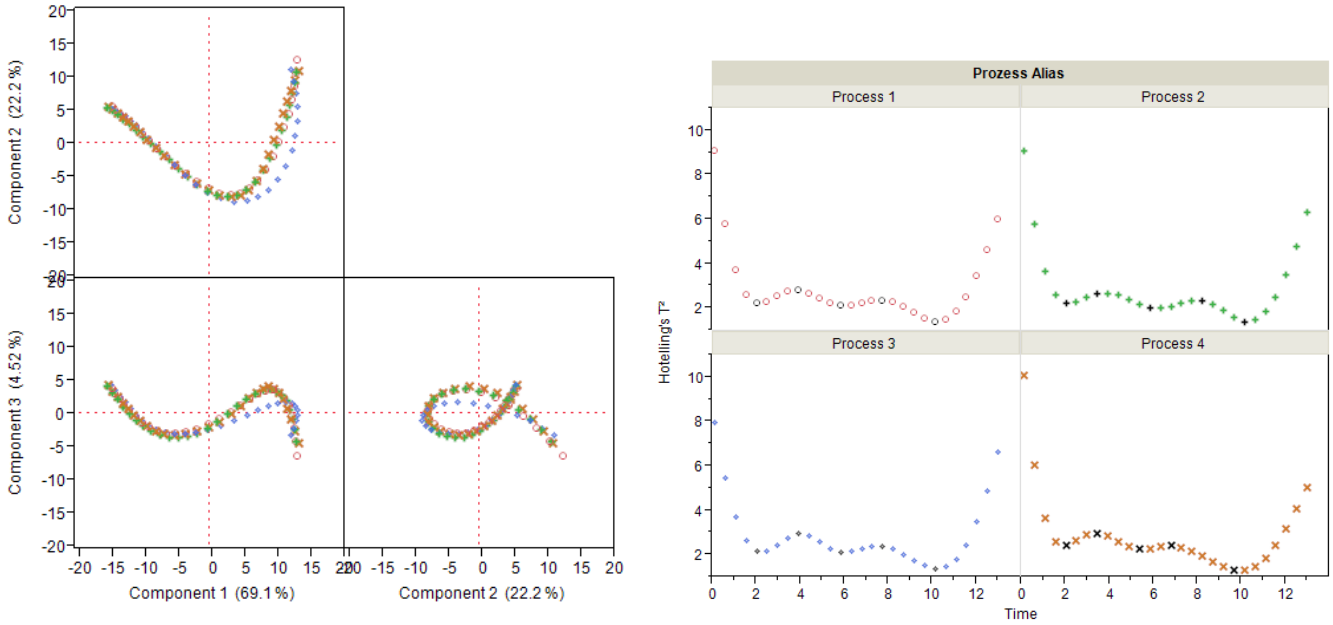

Figure 4 The left picture shows score plots of the first 3 principal components of the DPCA data set. The right picture shows a plot of time versus Hotelling's T^2 for all four fermentations. The black marked time points are considered to be phase shifts.

Although the automatic phase detection suggests several more phases, we here use the manual selection and choose only two phase shifts at time points 2 and 10. The phase in between roughly coinsides with the phase of exponential cell growth during fermentation (see Figure 2). Differences in the metabolite concentrations between different fermentations that come up during the exponential phase of cell growth are the most significant ones if we want to understand differences between fermentations. Differences that come up early in fermentation during the stationary phase of course also have a major impact on the fermentation results. But these can most certainly be explained with differences in the starting conditions of the fermentation and thus do not convey any further knowledge on the behaviour of the CHO cell culture during fermentation.

The module report now automatically offers a PCA analysis of the data to give the user a visual impression of the data set for the chosen phase. See Figure 5 for a 3D plot of our data mapped on the first three principal components. The time course is highlighted by moving from blue (day 4) to red (day 10). It is easy to see that, although we deal with replicates, one fermentation differs from the rest at early time points (diamond shaped) and one fermentation tends to move away from the rest at late time points (cross shaped).

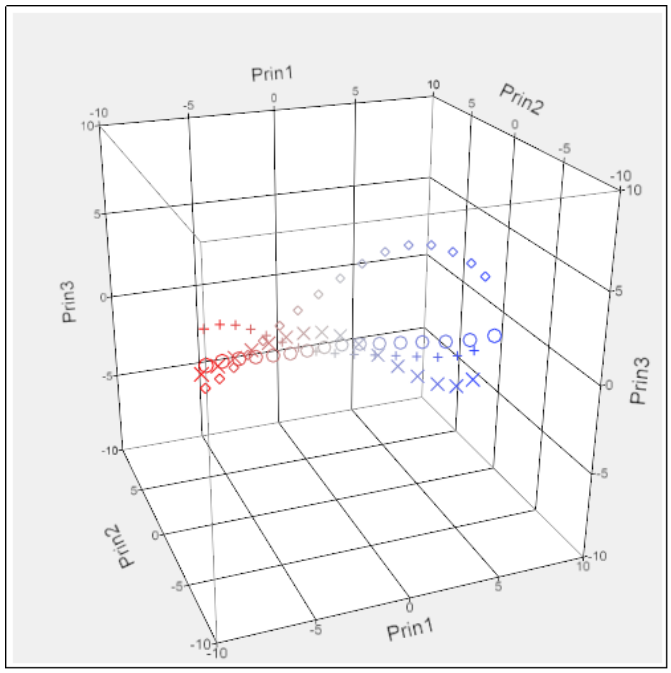

Figure 5 3D plot of the measurements from the exponential phase mapped on the first three principal components. The color coding blue to red indicates the time from day 2 to day 10.

The next step is to identify the metabolites responsible for these differences.

Step 3: Identification of key metabolites

The JMP module offers the possibility to assign the fermentations to different groups and then presents different alternatives to proceed. You can either select subsets of time points at which you want to identify the metabolites that explain your groupings or you can try to identify the metabolites that best explain your groupings over the whole time course. In our example we assign the diamond shaped fermentation to one group and the rest of the fermentations to another.

Next we choose to only use the first five time points for an analysis using JMP's own discriminant analysis method. The module generates a report which shows the initials F-ratios for every metabolite as a bar plot and three additional canonical. These plots are generated by choosing one of the three highest F-ratios. The module dialog furthermore offers the option to proceed with a stepwise variable selection if one variable is not sufficient to explain the differences. As Figure 6 shows, for our data set there are

several possible metabolites that can explain the difference between the two groups of fermentations.

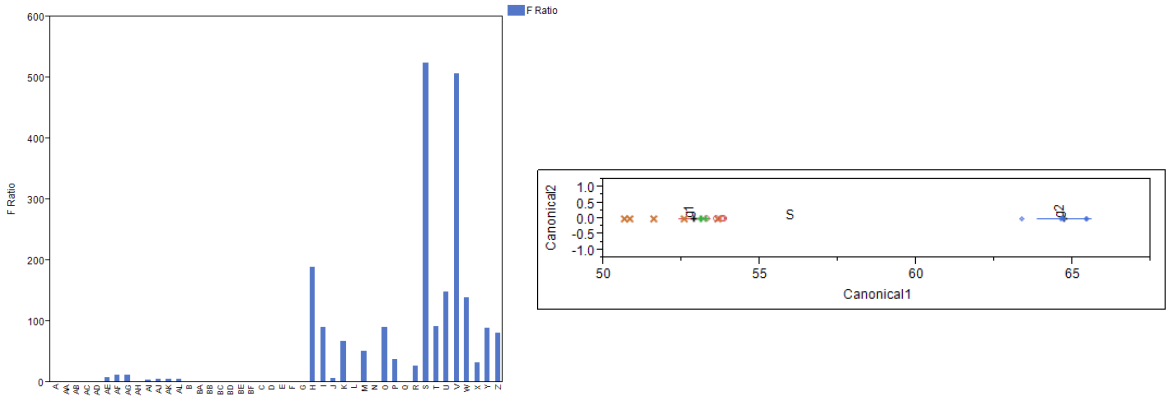

Figure 6 Results from the discriminant analysis. There are several metabolites with high F-ratios, which indicates that a good separation of both groups is possible. This can be seen for metabolite S in the right figure.

In a second approach we analyze the complete data set via the MEBA method. As a result of the MEBA we get a Hotelling's T^2 statistic for every metabolite. The higher the value of this statistic the more differs the metabolite between the two different groups of fermentation. The results for the example data set are comparable to the results of the discriminant analysis (Figure 7). Of course this is just a coincidence. If we look for differences between the cross-shaped fermentation and the rest, we do not find a good explanation for the differences with the discriminant method but at least one interesting metabolite via MEBA (data not shown).

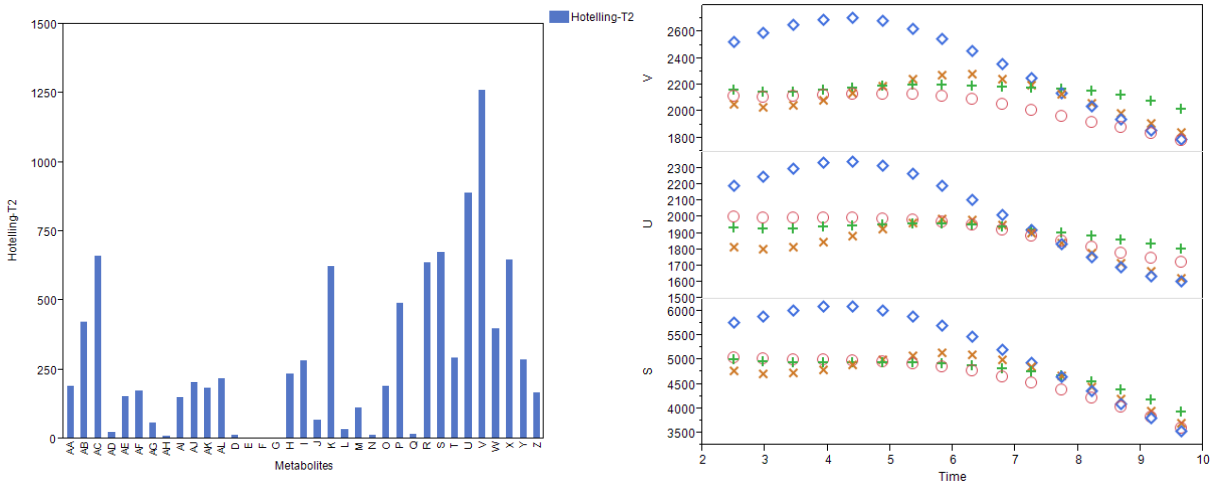

Figure 7 The MEBA method returns a statistic for every metabolite. The higher the value of this statistic the bigger the differences between the groups for a given metabolite. The right plot shows the top three discriminating metabolites.

For a last example we turn to the ASCA method. As this method is closely related to MANOVA it works best for balances experimental designs with enough replicates. Our example data set only offers four fermentations. We split the fermentations into two groups. The diamond and cross-shaped fermentations form group *g2*, the other two form group *g1*. We furthermore select one time point per day from the exponential phase data. After running the ASCA method, several different results are

returned. The first set of results highlights the principal components for the factor time, the factor groups and for the interaction between both factors. See Figure 8 and Figure 9 for plots of the different PCs. Additionally the most significant metabolites for the two factors and the interaction are returned. As Figure 10 shows, this is again metabolite V for the factor groups, which is due to its high impact on the diamond-shaped fermentation. This also highlights that there seems to be no metabolite that really discriminates our two chosen groups, which comes to no surprise when we think of the PCA plot. The factor time on the other hand has the most significant impact on metabolite D. There are no significant metabolites for the interaction between time and groups.

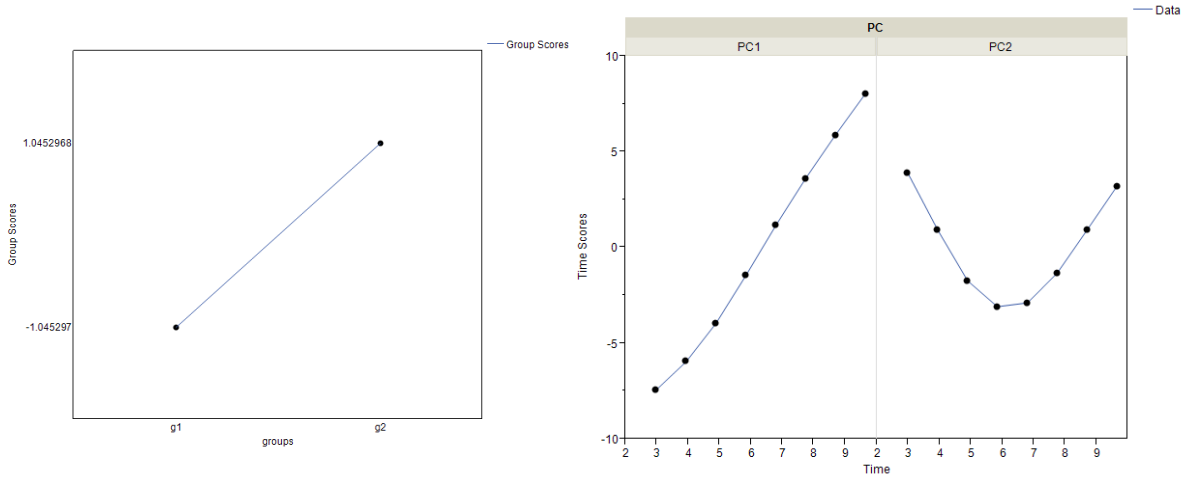

Figure 8 Principal component score plots for the factor groups (left) and time (right).

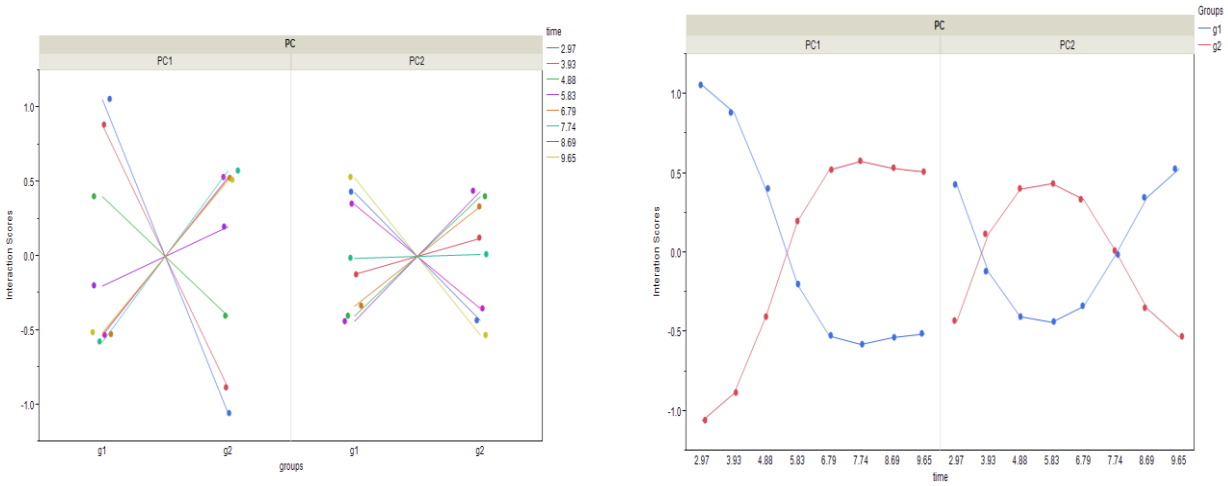

Figure 9 Principal components score plots for the interaction between the factors groups and time.

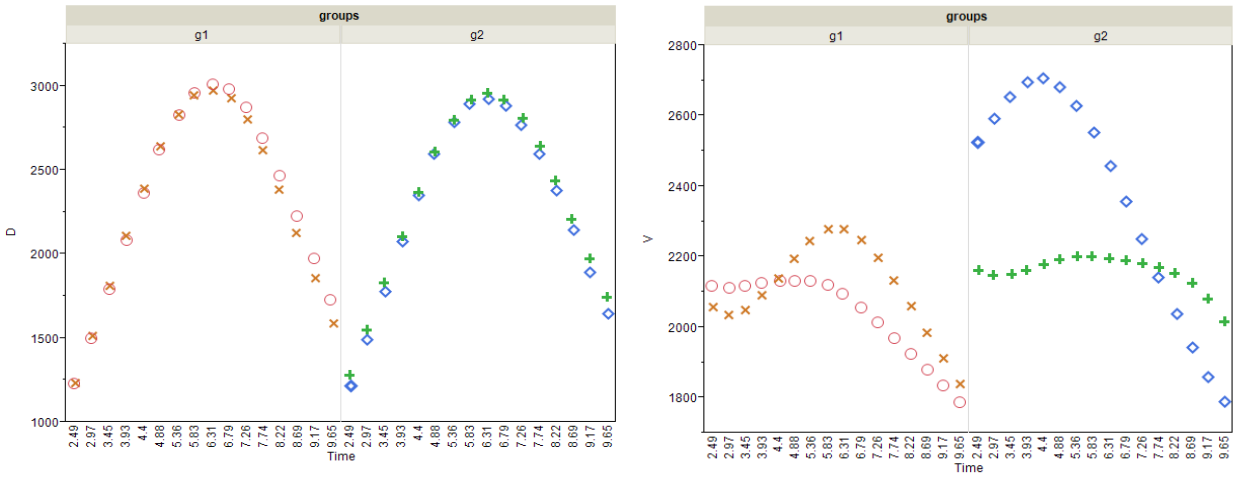

Figure 10 The most significant metabolite with respect to the factor time (left) and the factor groups (right). There were no significant metabolites for the interaction between both factors.

Summary

The new JMP module offers a platform for the analysis of metabolite data collected during fermentation of CHO cells. The flexibility and interactivity which JMP offers play key roles in the JMP module. It makes use of many method already present in JMP and adds additional methods by connecting to R. In this paper we concentrated mainly on showing the use of these additional methods. The application of methods like partitioning or neural networks to such data sets should be obvious. For example in lacking a good model of the CHO metabolism, the interaction profiler from the neural network platform can be used to help the user understand the impact of the different metabolites on viable cell density or production over the course of the fermentation.

The user can already follow some standard analysis workflows. For the future it is planned to heavily document every step of these workflows inside the module dialog to allow also non-statisticians to analyze metabolite data sets with JMP. As already mentioned before, the analysis methods and workflows of the consumption rates are still under construction. Furthermore, we plan to add metabolic flux analysis as an additional feature (see [4]) and evaluate the use of the module for other cell types such as E. coli.

References

[1] Zhigiang, A. Therapeutic monoclonal antibodies: from bench to clinic. Wiley & Sons. 2009

[2] Jayapal KP, Wlaschin KF, Hu W-S, Yap MGS. Recombinant protein therapeutics from CHO cells-20 years and counting. Chem Eng Prog. 2007;103:40–47.

[3] Birch JR, Racher AJ,. Antibody production, Advance drug delivery reviews 2006; 58; 671-685

[4] Ahn WS, Antoniewicz MR. Towards dynamic metabolic flux analysis in CHO cell cultures. Biotechnol J. 2012; 7(19);61-74

[5] Reinsel G. Elements of Multivariate time series analysis. Springer Series in Statistics

[6] Tai Y, Speed T. A multivariate empirical Bayes statistic for replicated microarray time course data. Annals of statistics. 2006; 34; 5; 2387-2412

[7] Vis et al. Statistical validation of megavariate effects in ASCA. BMC Bioinformatics. 2007; 8

[8] Doan et al. Detection of phase shifts in batch fermentation via statistical analysis of the online measurements: A case study with rifamycin B fermentation. J Biotechnol. 200,; 1332(2); 156-66

[9] Xia J, Sinelnikov I, Wihart D. MetATT: a web-based metabolomics tool for analyzing timer-series and two-factor data sets. Bioinformatics. 2011; 27 ; 2455-6

[10] Smilde et al. ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. Bioinformatics. 2005, 21; 3043-3048

[11] Giogino T. Computing and Visualizing Dynamic Time Warping Alignments in R: The dtw Package. Journal of Statistical Software. 2009;31(7); 1-24

[12] Schaub et al. Advancing Biopharmaceutical process development by system-level analysis and integration of omics data. Adv Biochem Eng Biotechnol. 2012;127;133-63