Data Mining to Investigate Relations Between Seed and Main Culture Performance

Dominik Driesch*, Michael Pfaff*, Reinhard Guthke
Hans Knöll Institute for Natural Product Research
Beutenbergstr. 11, D-07745 Jena, Germany
Phone: +49-3641-656820, Fax: +49-3641-656800
email: rguthke@pmail.hki-jena.de
*BioControl Jena GmbH
Wildenbruchstr. 15, D-07745 Jena, Germany
Phone: +49-3641-675511, Fax: +49-3641-675512
email: biocontrol@t-online.de

ABSTRACT: The performance of the seed culture is often crucial to the performance of the main culture. This statement does not only apply to modern biotechnological processes but also to more traditional biotechnological processes such as beer brewing. Therefore, a number of approaches have been made to describe the seed culture's vitality. Since some of them deal with rather complex data, validation of these methods is usually not trivial. We report here on a way to validate new methods for the estimation of yeast vitality in seed cultures with respect to the main cultures' performance.

KEYWORDS: Seed Culture, Main Culture, Beer Brewing, Yeast Vitality, Batch Process Analyzer, Decision Xpert

INTRODUCTION

Most of today's established models for growth of and product formation by micro-organisms, sophisticated as they might be, lack an important feature: they do not pay enough attention to the micro-organisms' history. These disadvantages become obvious for instance when trying to cope with the uncertainties of yeast management in breweries. Introducing vitality into modelling may bridge this gap between seed and main culture. Since a number of approaches have been made to describe yeast vitality, a method to validate these approaches with respect to their use in modelling is necessary. As vitality is somehow related to the lag-time of the main culture, any correlation that can be found between lag-time and vitality test signals gives proof of the test's applicability. Which main culture quality criterion is correlated to the vitality signals depends on the definition of vitality as well as on the process investigated. Lag-time plays an important role in the brewing process and is also highly desired to be controlable. Therefore lag-time was chosen as the quality criterion for the main cultures described in this work. Correlation between the signals and the quality criterion should be established by methods as simple as possible. However, signals can only be correlated by classical statistical methods as long as they are not complex. Otherwise, advanced statistical or data mining methods have to be applied.

SEED CULTURE VITALITY TEST

Saccharomyces carlsbergensis was cultivated continuously on wort containing 11.1° plato. To generate different seed culture states and different yeast vitalities, the cultivation was carried out under different cultivation parameters (temperature, dilution rate, oxygen supply). The vitality of cells of these seed cultures was tested prior to inoculation of the main cultures as described previously (Driesch 1998). Vitality signals of 16 seed cultures investigated in this work are shown in Fig. 1.

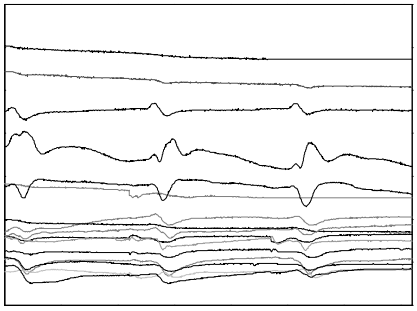

Figure 1: Normalized vitality signals of 16 yeast seed cultures taken prior to inoculation of main cultures

MAIN CULTURES AND MODELLING

The main cultures (i.e. beer fermentations in compliance with the German Reinheitsgebot) were carried out as a test system for the inoculum using the same wort as for the seed cultures. All main cultures were inoculated to yield a cell count (c_X) of $5 \cdot 10^6$ cells/ml. The temperature was 15°C for all main cultivations. The wort was saturated with sterile air at the beginning of the cultivations and no oxygen was supplied during the cultivation. The pH was left uncontrolled. Maltose, maltotriose, glucose, fructose and saccharose concentrations were measured by HPLC. These concentrations were added to yield the extract (c_S) which was taken into account for modelling. The ethanol concentration (c_E) was determined enzymatically (NADH related alcohol dehydrogenase reaction measured at 365 nm). According to literature the cultivation was assumed to be nitrogen (c_N) limited and the initial nitrogen concentration was set to 0.2 g/l. The cultivations were modelled by the following differential equation system which was fitted to the data of 16 fermentations.

$$m = \mu_{\max} \cdot \frac{c_S}{c_S + K_S} \cdot \frac{c_N}{c_N + K_N} \cdot (1 - e^{-t/t_L})$$

$$\frac{dc_X}{dt} = m \cdot c_X - k_E \cdot c_E \cdot c_X$$

$$\frac{dc_S}{dt} = -\frac{m \cdot c_X}{Y_{X/S}}$$

$$\frac{dc_E}{dt} = m \cdot c_X \cdot Y_{E/X}$$

$$\frac{dc_N}{dt} = -\frac{m \cdot c_X}{Y_{X/N}}$$

The model takes into account growth limitation with respect to nitrogen and sugars (extract) represented by the parameters K_N and K_S and cell starvation in the presence of ethanol represented by k_E. Since the yield coefficients Y and the parameters μ_{max}, K_N, K_S and k_E are strain dependent, they were fitted to the data of all runs (16 fermentations). The lag-time t_L quantifies the time the growth is delayed, indirectly representing the vitality of the seed culture in the main culture. This lag-time was assumed to be different for every single run. Consequently it was fitted to each main culture separately.

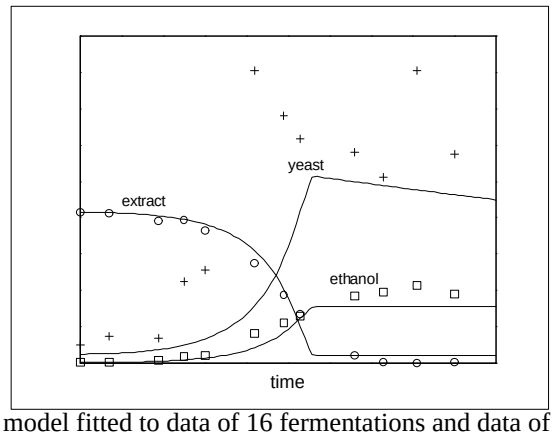

Figure 2: Response of the model fitted to data of 16 fermentations and data of one run (normalized units)

CORRELATIONS BETWEEN SEED CULTURE VITALITY AND MAIN CULTURE LAG-TIME

The higher the vitality of a seed culture the shorter the lag-time of the corresponding main culture should be. Therefore, a seed culture vitality test should produce a signal that correlates in one way or another with the lag-time of the main culture provided the vitality test is valid. We started with simple correlation tests. Fig. 3 (left) shows a plot of the mean value of the vitality signal versus the lag-time. No correlation could be found. This also applied when the range, i.e. the difference between the maximum and the minimum value of the vitality signal was plotted against the lag-time as shown in Fig. 3 (right).

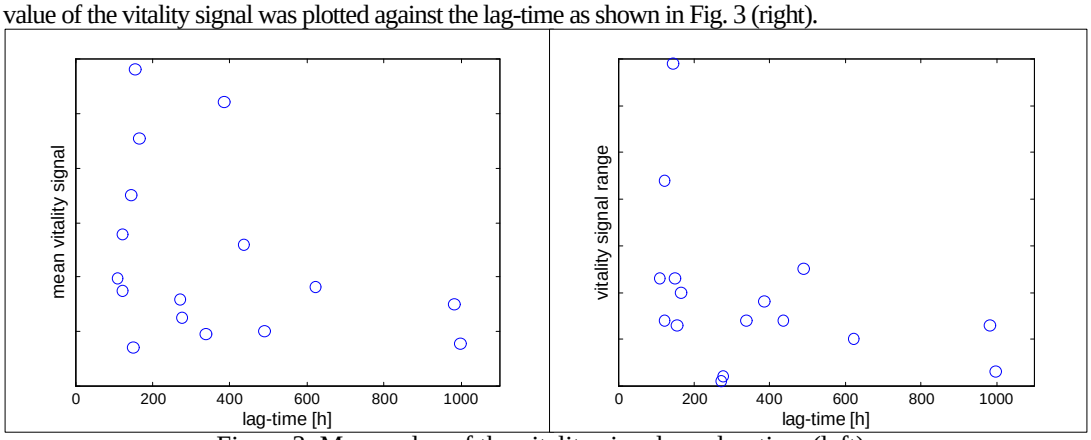

Figure 3: Mean value of the vitality signals vs. lag-time (left).
Range between maximum and minimum value of the vitality signals vs. lag-time (right)

As no correlation for any of the absolute or relative values of the vitality signals to the lag-times could be found it was suspected that the information might be more likely found in the shape of the signals and so we turned to wavelet analysis. Wavelets are known to analyse signals with respect to their shape at different scales. The vitality signals were analysed using biorthogonal spline wavelets. Examples for the best correlated wavelet coefficients are shown in Fig. 4. Again, no sufficient correlation to the lag-time was found.

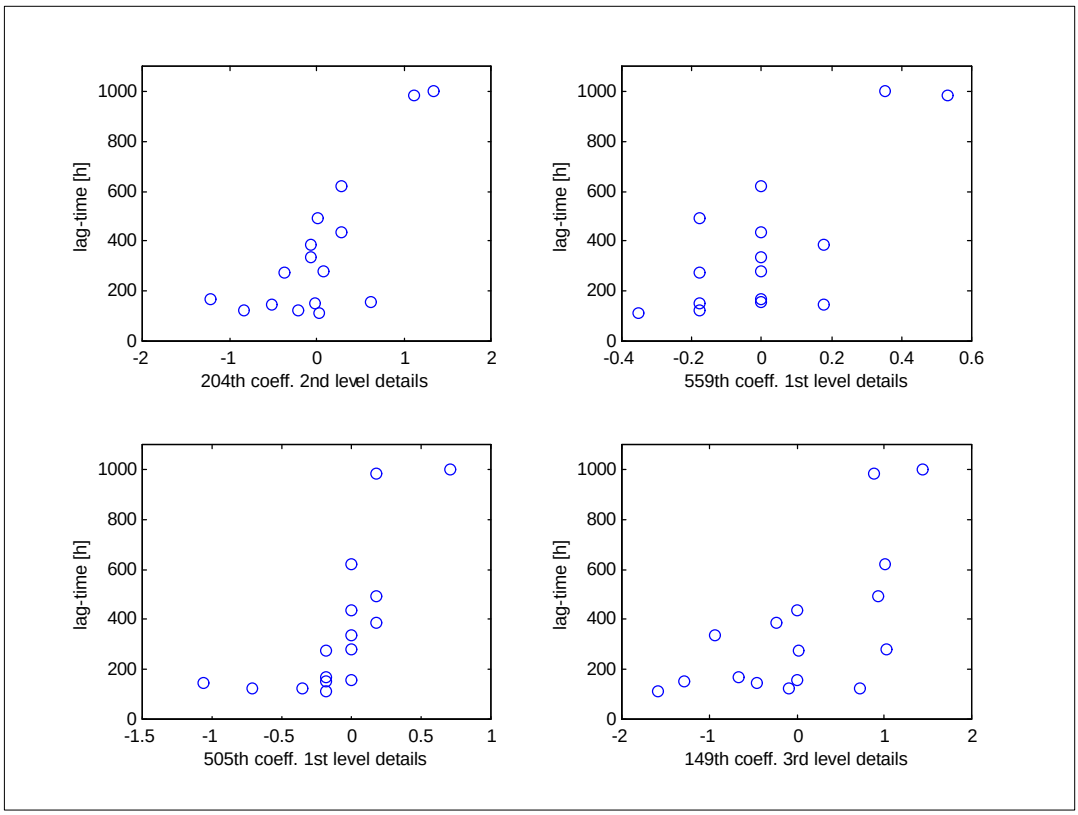

Figure 4: Lag-time vs. the four best correlating wavelet coefficients.
Signals were decomposed using biorthogonal spline wavelets.

We then subjected the vitality signals to data mining procedures using the ProcessAnalyzer (Locher 1996) and the DecisionXpert (Borgelt 1998), two Plug-Ins of the software tool DataEngine (MIT GmbH, Aachen, Germany). By combining these Plug-Ins DataEngine provides a method that can correlate pre-classified process outcomes (e.g. the main cultures classified by their lag-times) to complex process data by extracting decision rules out of the data. This method points to where the discriminating information is located inside the process data, if there is any at all. This method consists of three steps.

- Multi-scale filtering of the process data (e.g. vitality signals) using cubic spline wavelets. The information is split into details at different scales (from general approximation to detailed information).

- Trend decomposition of the filtered signals. Thereby the information concerning the shape of the signals is made accessible to algorithms that deal with absolute or relative values. The algorithm identifies extrema and inflection points in the signals and decomposes the signals into four elementary trend types. The signals are represented by a sequence of letters each of which denotes one of these elementary trends. The trends were matched between the signals and only those that fullfill certain matching criteria are further processed. The matched trends are described by qualitative parameters that concentrate the information for the next step.

- Generation (growing) of a decision tree to correlate the pre-classified process outcomes to the process data. The Plug-In DecisionXpert generates rules of the kind ‘IF variable_A of run x > B AND variable_C of run x < D THEN run x belongs to class E’. The rules have a tree-like structure, that means one decision node leads to the next decision node and finally to the leaves that hold the class memberships (Borgelt 1998).

Applying this strategy to the above described correlation problem leads to certain problems that arise especially when analysing noisy data. Decision trees can be grown for nearly every possible given class distribution when dealing with data sets that contain few runs and a high number of attributes. That means, correlations between two variables found by a decision tree do not necessarily represent a relation that really exists between these variabels.

Consequently the grown decision trees have to be validated themselves. To do so, the whole data set was classified with respect to the process outcome (lag-time). One run was taken off from each class (test data) and a decision tree was grown based on the remaining data set (training data). After that, the test data was subjected to the grown decision tree. It is important to select test data

sets that do not lay on class boundaries. The test data sets must then be classified correctly by the grown decision tree for a significant number of runs to give proof of an existing correlation between the process data and the process outcome (vitality signals and lag-time).

The data set was classified into two classes: class one contains runs with a lag-time shorter than 200 hours and class two all runs with lag-times longer than 200 hours. Runs 11, 12 and 13 of class one and runs 1, 4, 6 and 16 of class two were chosen as test data sets. This yields 12 possible combinations of test and training data sets. Twelve decision trees were grown and each was tested using the corresponding test data set.

Data Set	Lag-Time [h]	Class Membership
01	437	2
02	271	2
03	110	1
04	491	2
05	999	2
06	337	2
07	277	2
08	983	2
09	151	1
10	622	2
11	123	1
12	123	1
13	145	1
14	154	1
15	166	1
16	387	2

1. Data sets of seed culture vitality classified by the main culture's lag-time with their class memberships

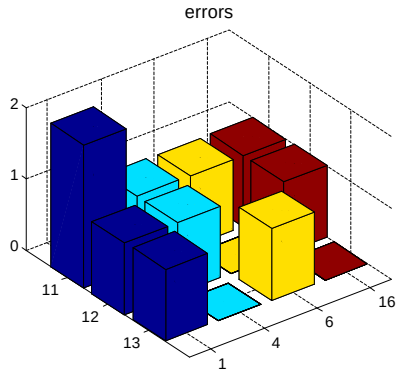

Figure 5: Errors made by the decision tree algorithm when classifying the test data sets by the grown decision trees. All possible test set combinations are displayed on the xy-plane and their errors on the z-axis

The errors that were made during classification of the test data sets are shown in Fig. 5. Ten errors did not significantly reduce the number of errors that would have been made by classifying all runs 'blind' into the same class (12 errors). Therefore it can be concluded, that the decision tree algorithm used has found the correlations by chance and not due to existing relations between the vitality signal of the seed cultures and the lag-time of the main cultures.

CONCLUSIONS

The DataEngine Plug-In DecisionXpert (MIT GmbH, Aachen, Germany) is capable of detecting complex correlations between two data sets even if they only exist incidentially for instance in noisy data. To avoid unintended detection of this kind of correlations the data has to be split into training and test data sets and the extracted rules have to be verified by the test data sets.

Nonetheless, no correlation between the yeast vitality of the seed cultures and the lag-time of the corresponding main cultures could be found. The vitality test method applied to the seed cultures is therefore not capable of determining the vitality of yeast cells with respect to the main culture performance criterion investigated (lag-time).

REFERENCES

Borgelt, C. (1998): A Decision Tree Plug-In for DataEngine. 6th European Congress on Intelligent Techniques & Soft Computing Aachen, Germany, September 7-10, 1998, Proceedings, pp. 1299-1303.

Driesch, D., Pfaff, M. (1998): Bioactivity Pattern Recognition. 6th European Congress on Intelligent Techniques & Soft Computing Aachen, Germany, September 7-10, 1998, Proceedings, pp. 1574-1579.

Locher, G., Bakshi, B., V Stephanopoulos, G., Stephanopoulos, G., Schügerl, K. (1996): Ein Ansatz zur automatischen Umwandlung von Rohdaten in Regeln. Teil 1+2, Automatisierungstechnik 44, Nummer 2+3, pp. 61-70 + 138-145.