CHIMIA **2005**, *59*, No. 10

Chimia 59 (2005) 753–755

ISSN 0009–4293

Process Optimisation Based on Large Databases of Routinely Monitored Industrial Process Data

Karin Kovar[a]*, Thomas K. Friedli[a,b], Dusan Roubicek[c], David S. Langenegger[a], Markus Keller[a], and Hans-Peter Meyer[d]

Abstract: Huge amounts of data are routinely logged and stored during the monitoring of biotechnological production processes. A concept is described to extract and analyse the information these data contain and to subsequently apply it for process improvement. In total, roughly 100,000 time series of raw and derived signals which stemmed from 173 high-cell-density processes with recombinant microorganisms at 50 m^3 scale (working volume) were processed. As is often the case, no mathematical process models were readily available and therefore data-driven, computer-intensive methods were applied. These endeavours helped to stimulate a change in manufacturing strategy, which in turn has led to an increase in the final product titre of 26% on average.

Keywords: Computer-intensive methods · Data-driven statistical methods · Intervention-impact analysis · Large database

1. Introduction

The use of statistical methodology to support experimental work during the development of a biotechnological process is well accepted (*e.g.* strain screening, exploration of relationships). However, once the particular process has passed into the production stage, it is usually fixed and not further improved. In order to master the demanding online process control while simultaneously complying with the requirements of the regulatory agencies, huge amounts of data are typically logged and stored during routine (cGMP) monitoring of production process. Although the information these data contain could be advantageously used to gain increased understanding of the process and thereby used for its optimisation, these data are not usually processed further and the information available therein is frequently wasted.

In this paper, a concept and methods for activating such unproductive data are presented. Improvements at the stages of data management, data analysis, laboratory experiments and process control are included [1]. The aims of this generalised strategy are:

(i) the foundation of a homogenous database of historical processes which can be compared with the actual running process using the existing process control system (PCS);
(ii) the extraction of available process knowledge and a proposal for suitable experimental design;
(iii) improvement of process productivity; and
(iv) improvement of process quality.

Ideally, when the historical database contains sufficient variability, influence factors can be extracted and changes to the manufacturing strategy directly introduced on production scale, thus avoiding demanding laboratory experimentation (Fig. 1). Generally speaking, it is not impossible to introduce changes to cGMP performed processes, but these are subject to strict regulations and therefore have to be justified by a reliable intervention-impact analysis.

2. Material and Methods

The database evaluated contained details of 173 fedbatch processes with 90 variables logged both online and offline per process. After preprocessing the raw data and computing derived variables, the whole database consisted of roughly 100,000 time series. Data logged within the actual run were processed in parallel using the *Lucullus* Process Information Management System (Biospectra AG, Schlieren CH).

The software enabling automated data evaluation comprises (i) the extraction of time series of meaningful signals including, for example, brushing, polishing (denoising), data synchronising on common time grid and smoothing, derivations, integrations and arithmetical calculations, (ii) the extraction of key characteristics/values (*e.g.* final product titre, maximum productivity) and (iii) a process comparison including the rating of the processes.

In the Table the following methods are listed: the standard statistical methods which are described in the literature [2], and a statistical software implementation

*Correspondence: Prof. Dr. K. Kovar[a]
Tel.: +41 44 789 9733
Fax: +41 44 789 9950
E-Mail: k.kovar@hsw.ch
www.bioprocess.ch and *www.biotechLAB.ch*
[a]University of Applied Sciences Zurich
Postfach 335
CH-8820 Waedenswil, Switzerland
[b]University of Bern
Institute of Mathematical Statistics and Actuarial Science (IMSV)
Sidlerstrasse 5
CH-3012 Bern, Switzerland
[c]Lonza biotech s.r.o.
Okruzni 134
CZ-281 61 Kourim, Czech Republic
[d]Lonza AG
Walliser Werke
Postfach
CH-3930 Visp, Switzerland

Table 1. List of (statistical) methods implemented

method	implementation	references
robust local regression	Loess regression	[3]
robust filtering	median filtering, waveshrink	[4][5]
nonparametric regression	smoothing spline	[6]
robust local numerical derivation	numerical derivation of Loess filtered and spline interpolated data	[7]
robust local numerical integration	numerical integration of Loess filtered and spline interpolated data	[7]
confidence intervals	Bootstrap simulation	[8]
univariate statistics	minimum, maximum, arithmetic mean, median, variance, standard deviation (s), standard error, MAD, interquartils range (IQR), s_{IQR}, s_{MAD}	[9]
rating	ranking with stochastical breakink ties	[9]

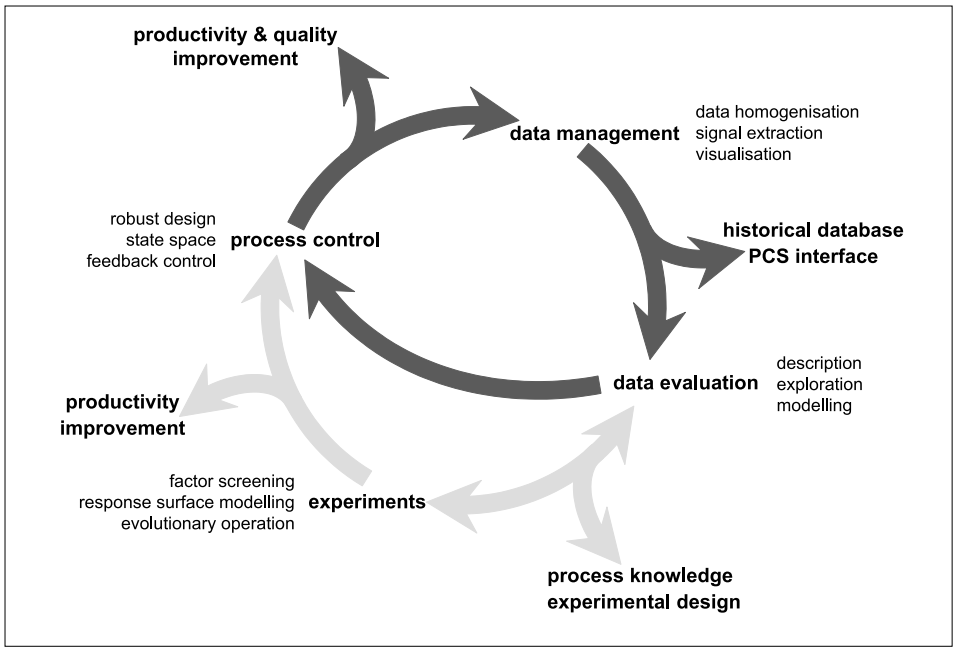

Fig. 1. General framework for the implementation of statistical data analysis in a biotechnological production facility

of our own which was developed using S-Plus statistical software/language (Insightful Inc.) and, as remains to be seen, may be transferred into open source R-code. In addition, several different types of conventional as well as specialised univariate and bivariate data visualisation were automated (*e.g.* overlay line-plots and overlay scatterplots, Boxplots, residual plots, scatter matrix plots).

3. Results

Based on the statistical analysis of data from 173 high-cell-density processes, an investigation was made of the reasonable possibilities of improving the production process which can be drawn from the analysis of a database with historical process data. The monitored variables were first preprocessed using non-parametric regression models such as splines and wavelets. Overlay and conditional plots were used to identify the best performing processes with respect to several criteria set (*e.g.* for the time-course of product concentration shown in Fig. 2). Then the processes were grouped on the basis of *a priori* knowledge available (*e.g.* four feed profiles, three reactor types, and, possibly, various seed cultures and strains). Finally, five process groups were identified and the variation with respect to the final product concentration was illustrated with Boxplots (Fig. 3). Since the difference between the first and second groups is not statistically significant and the fifth group contains processes where gross failures in process control were discovered, only three process groups were actually identified. These results were also confirmed by cluster analysis, which does not make use of any *a priori* information. Based on the variance of the groups, process capability (*i.e.* the potential for process improvement) was identified as approximately 5% improvement potential from better process control and roughly 32% from the adaptation of culture conditions and/or the time course of feed addition. These findings stimulated substantial changes in manufacturing strategy, which in turn have led to an increase in the final product titre of 26% on average (data not shown).

To plan the number of experiments to be performed in production-scale, an intervention-impact analysis has been carried out. This was based on the analysis of variance of the historical (and then available) data and predicts the further process performance after certain changes in the manufacturing strategy (Fig. 4). The aim of this method is firstly to determine the number of process runs needed to verify an expected process improvement and, secondly, to determine the minimal process improvement which can be verified with a given number of process runs. Finally, fewer than five process runs were needed to detect the effective improvement in the final product titre (26% on average) as being a statistically significant event.

4. Conclusions

Continuous optimisation of industrial production processes in biotechnology is a controversial topic. A common prejudice claims that, under cGMP conditions, there is insufficient natural variability in the process data to identify the influencing factors and relationships needed to effect process improvement. In our experience, as detailed in this paper, this is generally not true.

As a basis for management decision-making, it is worthwhile analysing these data and performing intervention-impact analyses. However, explorative data analysis or even a simple visualisation of such a huge database cannot be performed 'manually' with ordinary spreadsheet tools such as Microsoft Excel. Thus, in the scope of the project, generic methodology for (automated) handling and (automated) statistical evaluation of large databases of biotechnological process data were developed. The data-driven

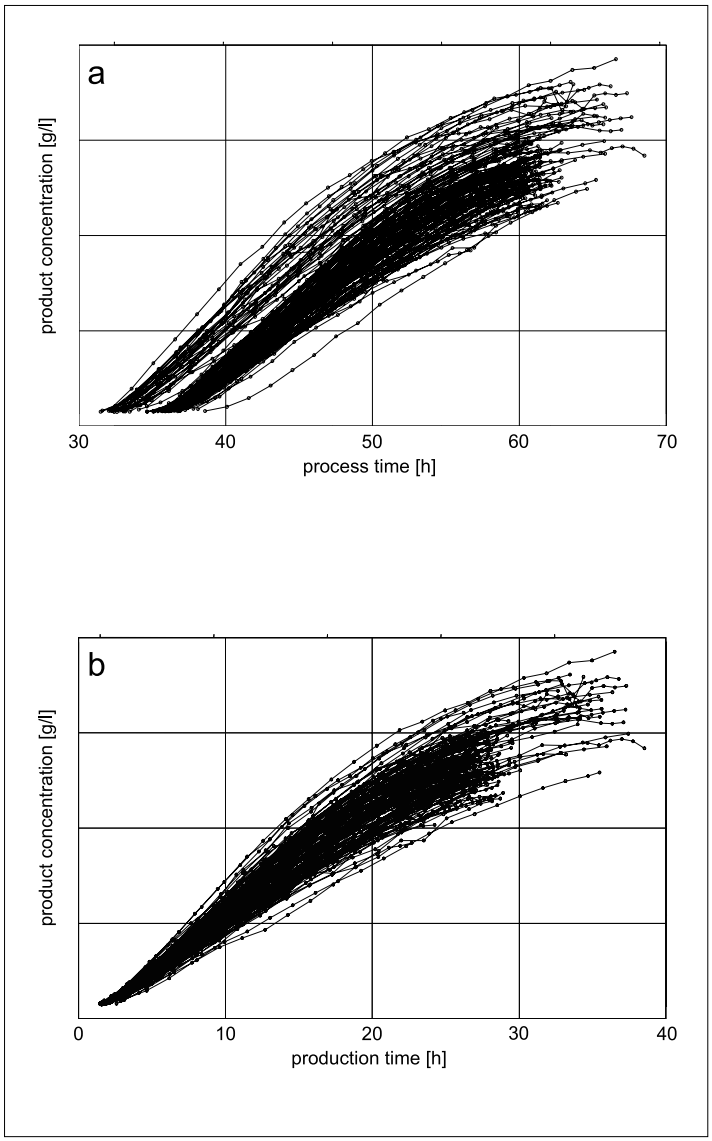

Fig. 2. Overlay plot visualising the time course of product concentration of 173 processes. a) Brushed raw data (*i.e.* data without outliers) plotted on the original time axis with process inoculation by time 0 hours; b) brushed raw data synchronised for time 0 h equivalent to the beginning of the production phase (*i.e.* induction of the gene expression)

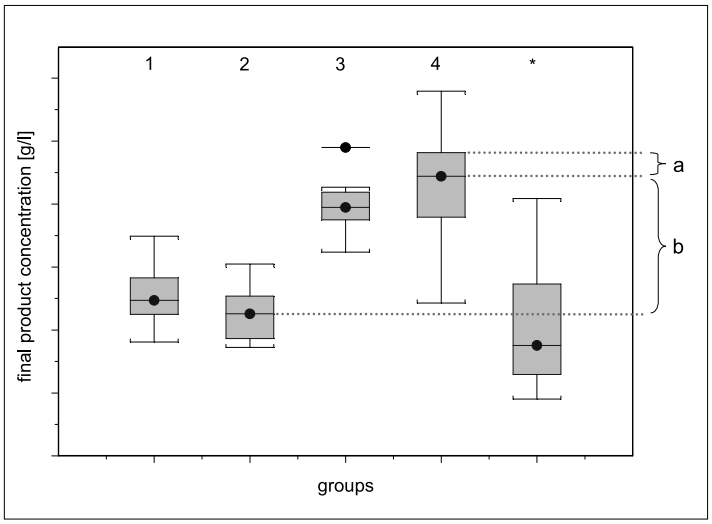

Fig. 3. Boxplots of product concentration at the time of harvesting for five groups of processes. The boxes contain 50% of the values; the black dots in these boxes denote the median values. Groups marked with an asterisk (*) show data from processes where problems in process control were clearly identified. The difference between groups 1 and 2 is not statistically significant. a) 5.2% Improvement potential from better process control; b) 31.9% improvement potential from adaptation of culture conditions and/or feed strategy.

process models developed can further serve as (i) 'software sensors' to monitor those variables for which commercial online sensors do not exist, and (ii) as an early-warning-system to detect possible abnormalities during process runs.

Acknowledgement

The work was funded by CTI 5574.2 FHS project as well as by Biospectra AG (Schlieren, CH), Lonza AG (Visp, CH) and Lonza s.r.o. (Kourim, CZ).

Received: August 2, 2005

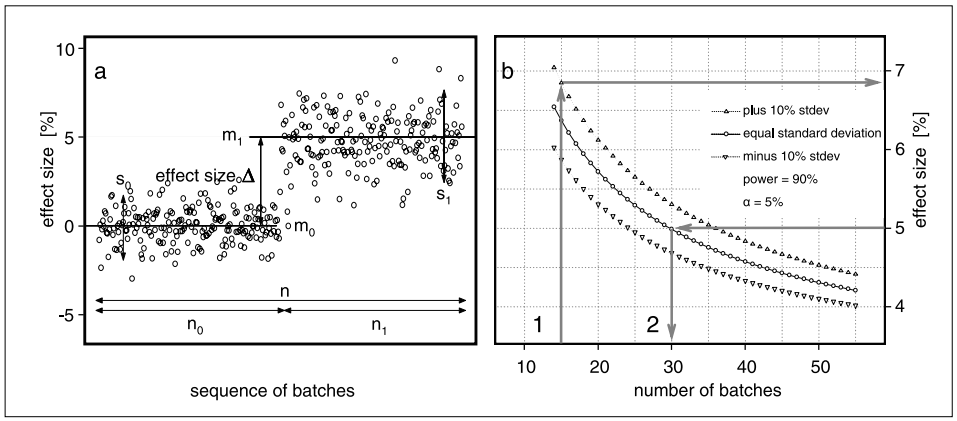

Fig. 4. Intervention-impact analysis. a) Definition of a process improvement using an effect size (n_0 and n_1 = number of processes before and after intervention/change respectively; m_0, m_1 and s_0, s_1,= means and standard deviations of the two data sets); b) number of process runs (*i.e.* experiments or observations) *vs.* effect size. The arrows indicate the following use-cases: 1) with a total of 15 batches an improvement of 6.75% is detected when the standard deviation of the new processes is simultaneously 10% lower than the standard deviation of previous processes; 2) in total, 30 processes are needed to achieve an improvement of 5% when the standard deviations of both means, *i.e.* before and after, are equal.)

[1] K. Kovar, T. Friedli, D. Langenegger, M. Keller. Project report, CTI 5541.2 FHS, **2004**. *http://www.bbt.admin.ch/kti/d/*

[2] T. Hastie, R. Tibshirani, J. Friedman, 'The Elements of Statistical Learning: Data Mining, Inference and Prediction', Springer, **2001**.

[3] W.S. Cleveland, *J. Am. Stat. Assoc.* **1979**, *74*, 829.

[4] J.W. Tukey, 'Exploratory Data Analysis', Addison-Wesley, Reading MA, **1977**.

[5] D.L. Donoho, I.M. Johnstone, *J. Am. Stat. Assoc.* **1995**, *90*, 1200.

[6] R.L. Eubank, 'Smoothing Splines and Nonparametric Regression', Marcel Dekker, New York and Basel, **1988**.

[7] T.K. Friedli, proprietary development, **2004**.

[8] B. Efron, R. Tibshirani, *Stat. Sci.* **1986**, *1*, 54.

[9] W.N. Venables, B.D. Ripley, 'Modern Applied Statistics with S', Springer, New York, USA, **2002**.