Process Biochemistry 49 (2014) 188–194

Contents lists available at ScienceDirect

Process Biochemistry

journal homepage: www.elsevier.com/locate/procbio

Prediction of product formation in 2-keto-L-gulonic acid fermentation through Bayesian combination of multiple neural networks

Tao Wang [a], Junwei Sun [b], Weidong Zhang [a], Jingqi Yuan [a,*]

[a] Department of Automation, Shanghai Jiao Tong University, and the Key Laboratory of System Control and Information Processing, Ministry of Education of China, 800 Dongchuan Road, 200240 Shanghai, China
[b] HeBei Welcome Pharmaceutical Co. Ltd., 11 North Zhaiying Street, 050031 Shijiazhuang, China

ARTICLE INFO

Article history:
Received 14 July 2013
Received in revised form 30 October 2013
Accepted 5 November 2013
Available online 14 November 2013

Keywords:
2-Keto-L-gulonic acid
Mixed culture
Batch classification
Product formation
Bayesian combined neural networks

ABSTRACT

As the key precursor for L-ascorbic acid synthesis, 2-keto-L-gulonic acid (2-KGA) is widely produced by the mixed culture of *Bacillus megaterium* and *Ketogulonicigenium vulgare*. In this study, a Bayesian combination of multiple neural networks is developed to obtain accurate prediction of the product formation. The historical batches are classified into three categories with a batch classification algorithm based on the statistical analysis of the product formation profiles. For each category, an artificial neural network is constructed. The input vector of the neural network consists of a series of time-discretized process variables. The output of the neural network is the predicted product formation. The training database for each neural network is composed of both the input–output data pairs from the historical bathes in the corresponding category, and all the available data pairs collected from the batch of present interest. The prediction of the product formation is practiced through a Bayesian combination of three trained neural networks. Validation was carried out in a Chinese pharmaceutical factory for 140 industrial batches in total, and the average root mean square error (*RMSE*) is 2.2% and 2.6% for 4 h and 8 h ahead prediction of product formation, respectively.

1. Introduction

L-Ascorbic acid, also known as vitamin C, is an essential nutrient for human beings and certain other animal species. Due to its indispensable physiological function, it is widely used in pharmaceutical formulations, food, beverage and animal feed. The industrial synthesis of L-ascorbic acid has two main routes. The Reichstein process discovered in 1930s applies a single pre-fermentation followed by a purely chemical process [1]. The modern two-step fermentation process is originally developed in China in 1970s [2]. In the first fermentation stage, L-sorbose is produced from D-sorbitol by batch culture of *Acetobacter melanogenum* with a high molar yield of about 96%. In the second stage, mixed culture of *Ketogulonicigenium vulgare* (*K. vulgare*) and *Bacillus megaterium* (*B. megaterium*) is practiced to produce 2-keto-L-gulonic acid (2-KGA) with L-sorbose as substrate [3,4]. 2-KGA is then converted to L-ascorbic acid by catalytic reactions.

Process monitoring and optimization have been paid great attention motivated by the increasing pressure to improve process efficiency with reduced operation costs in pharmaceutical industry. For instance, the prediction of product formation is of vital importance both for the production scheduling and early detection of faulty batches [5–7], so that the total profit of the fermentation workshop could be optimized. The purpose of this work is to develop an on-line application oriented predictor for the product formation in 2-KGA fermentation, which is going to be utilized in process performance evaluating and optimal scheduling.

Artificial neural networks (ANN), as a class of universal approximator [8], have been widely applied in process monitoring and optimization [9–11] because of its powerful input–output data mapping ability for nonlinear systems. Neural networks ensemble, which has been proposed for improving the performance and training of neural networks [12], is a collection of finite number of neural networks or other types of predictors that trained for the same task. Both theoretical and empirical studies have demonstrated that a good ensemble is one where the individual classifiers in the ensemble are accurate, and it could make the prediction errors on different parts of the input space [13–15].

In recent years, Bayesian combined predictor (BCP), which is derived from Bayesian analysis [16], has become an important approach for neural networks ensemble [17]. BCP is presented as a probabilistically motivated predictor for time series prediction. By incorporating an appropriate number of specialized predictors, the combination module can activate the appropriate predictor at the appropriate time, and therefore yield superior performance [18–23].

* Corresponding author. Tel.: +86 21 34204055; fax: +86 21 34204055.
E-mail addresses: wangtaovc@163.com (T. Wang), jqyuan@sjtu.edu.cn (J. Yuan).

1359-5113/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.procbio.2013.11.003

Nomenclature

Alk	alkali consumption (kg)
$C_{Alk,R}$	concentration of alkali solution (kg m^{-3})
C_P	product concentration in the medium (kg m^{-3})
C	cost parameter of SVM
e	learning error
F_{Alk}	feeding rate of alkali solution (m^3 h^{-1})
$K(X_i,X)$	Kernel function of SVM
L	medium level (m)
m	the discretization step length which equals T_D/τ
n_r	the number of historical batches being classified into the rth category after terminating
nh	the number of all the historical batches
N_i	number of the input–output data pairs of the ith batch
N_{n+1}	number of the available input–output data pairs belonging to the $(n+1)$th batch
P	product formation (kg)
$\hat{P}$	predicted product formation (kg)
$\hat{P}_{BCP}$	predicted product formation through Bayesian combined predictor (kg)
$\hat{P}_r(T_k+T_P)$	predicted product formation of rth ANN predictor for the time T_k+T_P (kg)
P_{ini}	initial product in the fermenter (kg)
P_m	statistical mean of historical product formation (kg)
$p^r_{T_k}$	the posterior probability of choosing rth ANN as the predictor at T_k
$RMSE$	root mean square error
S_F	cross-sectional area of the fermenter (m^2)
SU_i, SD_i	classification variables of the ith batch
t_k	sampling time (h)
$t_{n+1,f}$	current cultivation time of the $(n+1)$th batch (h)
T_D	width of the input data window (h)
T_k	cultivation time at the right border of kth input data window, especially T_1 is the start time of the prediction of product formation (h)
T_S	the sampling interval for off-line assays (h)
T_M	moving step length of data windows (h)
T_P	width of the prediction window (h)
x	vector of discretized process variables
X	input vector of the neural network
Y	output vector of the neural network
Z	a stochastic process which is defined as follows: $Z=n$ if at time T_k the nth ANN predictor is chosen
σ	standard deviation of normalized product formation
τ	discretization time interval of the input data window (h)
ε	classification precision in SVM
$\theta_{1\sim n}$	the collective input–output data pairs generated from n historical batches belonging to the corresponding category
θ_{n+1}	the collective input–output data pairs for the $(n+1)$th batch
λ	confidence coefficients corresponding to 60% confidence limits

In this study, the prediction of product formation is achieved by Bayesian combination of three neural networks. Each neural network is featured with its separate training database. First, the historical batches are evaluated according to product formation performance, and classified into three categories: excellent, normal and poor. Then, for each category, an artificial neural network is constructed. The input–output data pairs for the neural network training are generated with regularly measured variables and sampling assays. Specifically, the past accumulated product formation is included in the input data. The training database for each neural network is composed of both the input–output data pairs from the historical bathes in the corresponding category, and all the available data pairs collected from the batch of present interest. Finally, Bayesian combination of three neural networks is utilized to perform the 4 h and 8 h ahead prediction of product formation. Once the batch of present interest is terminated, its data will be added into the corresponding category of historical batches according to its batch classification result. The effectiveness of the proposed method is demonstrated with industrial fermentation data. Also, a comparison with SVM based rolling learning-prediction approach in the literature is presented.

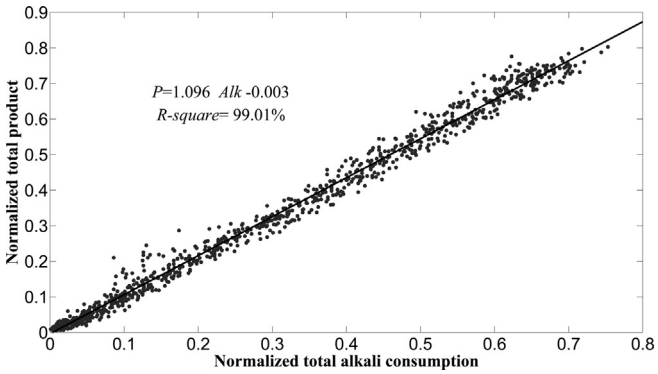

Fig. 1. Regression analysis between the accumulated alkali consumption and product formation.

2. Materials and methods

2.1. Process description

The 2-KGA mixed cultivation is carried out at 30 °C in an air-lift bioreactor with a volume of 200 m^3 in a Chinese pharmaceutical factory. The standard recipe of the initial medium contains 8.0% L-sorbose, 1.0% corn steep liquor, 0.2% carbamide, 0.1% KH_2PO_4, 0.01% $MgSO_4{\cdot}7H_2O$. All media are autoclaved before using. During the fermentation, alkali solution is automatically fed to control the pH value at about 7.0. To avoid substrate inhibition, the initial batch medium contains only about half of the total substrate, while the rest L-sorbose is fed into the bioreactor in the middle phase of fermentation.

Process variables such as temperature, pH, aeration rate, liquid level are on-line measured by the Distributed Control System (DCS). The 2-KGA concentration is off-line assayed regularly by iodometry titration [24], with a sampling interval of 4 h. Similarly, L-sorbose concentration is determined off-line by anthrone colorimetric method [25]. Limited sampling assays of L-sorbose concentration are available close to the end of cultivation, which are only used to determine the termination time. Typically, a batch will be terminated when the residual L-sorbose concentration becomes less than 1 kg m^{-3}. The microscopic examination of sample is regularly carried out to detect microbial contamination, but the biomass concentration is usually not measured in the mixed cultivation.

2.2. Generation of input–output data pairs for ANN training

During the mixed culture, the alkali solution is fed continuously so that pH of the medium is controlled at *c.a.* 7.0. On the other hand, the accumulated alkali consumption *Alk* is found to be approximately proportional to the accumulated product formation *P*. A linear regression analysis with the data of industrial batches demonstrates the close relationship between these two process variables, as shown in Fig. 1, where the coefficient of determination (*R-square*) is 99.01%. Therefore, besides the accumulated L-sorbose consumption and product formation, the accumulated alkali consumption can be used as an input variable of ANN.

The typical profiles of these accumulated variables are shown in Fig. 2, where the discrete time interval of 4 h is chosen, the same as the sampling interval for off-line assays T_S. These accumulated process variables are closely associated to future product formation.

However, in 2-KGA industrial fermentation, L-sorbose consumption and biomass concentration, despite of their importance, are not used for neural networks training for lack of routinely available data. Therefore, only the accumulated product formation and alkali consumption are used as the input variables of ANN. These two