International Journal of Biological, Life and Agricultural Sciences
ISSN: 2415-6612
Vol:2, No:1, 2008

# Predictive Clustering Hybrid Regression (pCHR) Approach and Its Application to Sucrose-Based Biohydrogen Production

Nikhil, Ari Visa, Chin-Chao Chen, Chiu-Yue Lin, Jaakko A. Puhakka, and Olli Yli-Harja

***Abstract*—**A predictive clustering hybrid regression (pCHR) approach was developed and evaluated using dataset from $H_2$-producing sucrose-based bioreactor operated for 15 months. The aim was to model and predict the $H_2$-production rate using information available about envirome and metabolome of the bioprocess. Self-organizing maps (SOM) and Sammon map were used to visualize the dataset and to identify main metabolic patterns and clusters in bioprocess data. Three metabolic clusters: acetate coupled with other metabolites, butyrate only, and transition phases were detected. The developed pCHR model combines principles of k-means clustering, kNN classification and regression techniques. The model performed well in modeling and predicting the $H_2$-production rate with mean square error values of 0.0014 and 0.0032, respectively.

***Keywords*—**Biohydrogen, bioprocess modeling, clustering hybrid regression.

## I. Introduction

MODERN systems biology approaches [1]–[3] have been successfully applied in the fields of pharmaceuticals and health sciences. The complexity of bioprocesses involved in environmental biotechnology and fermentation processes throws a new challenge [4], [5]. One of such bioprocesses is anaerobic digestion of sugar (usually glucose and sucrose) to produce $H_2$ as an energy carrier [6].

The fossil fuel dependency of energy economy today results in global warming, air pollution and environmental and health problems. Hydrogen ($H_2$) produced from renewable energy sources offers a clean alternative for the fossil fuels [7], [8]. Today, several techniques for sustainable $H_2$-production exist including microbiological fermentation processes [9].

Microbiological dark fermentation, involving mixed microbial cultures, can be used to produce $H_2$ from biomass or organic waste materials [10], [11]. $H_2$ production is an intermediate step in the anaerobic degradation of organic material, where $H_2$ and $CO_2$, and organic acids and alcohols are the end-products of the bioprocess [12], [13]. The fermentative biohydrogen research aims at manipulating and guiding the anaerobic digestion process in such a way that $H_2$ is produced as a major product. It is achieved by inhibiting the hydrogen consumers and methane producers; and controlling the various physical, chemical, biochemical and biological parameters [14].

In the growing ome-onyms, [15] introduces the term envirome for the state vector containing all relevant extra-cellular variables, e.g. concentrations of chemical compounds, control and operational physical variables. The interactions between envirome and other omes can be schematically represented as in Fig. 1. The biochemical information flow from genome to metabolome is shown as solid arrow. Other possible interactions are represented as dashed arrows. The metabolome interacts with the envirome via excretion or uptake of metabolites. The metabolome does not define the envirome in the same manner as the genome defines the transcriptome. The dashed arrow is used to indicate this difference.

The successful application of systems biology approach to biohydrogen process requires optimizing the envirome such that high and sustainable $H_2$-production rates are achieved [16]–[18]. The optimal conditions at envirome level are desired so that $H_2$-production is maximal and is also sustainable. During this stage, metabolome data is also analyzed for better understanding and control of bioreactor. Once the optimal envirome conditions are known for a desired metabolic profile with high $H_2$-production rates, more –omics related information could be included to better optimize, control and engineer the bioprocess.

In modeling and control of complex systems such as biotechnological processes, it is usually assumed that a global, analytical system model can be defined [19]. The kinetic and

Manuscript received on June 24, 2008.

Nikhil is with Department of Signal Processing; and Department of Chemistry and Bioengineering, Tampere University of Technology, Tampere 33720, Finland (phone: 358-3-3115-4956; fax: 358-3-3115-4989; e-mail: nikhil@tut.fi).

Ari Visa is with Department of Signal Processing, Tampere University of Technology, Finland (ari.visa@tut.fi).

Chin-Chao Chen is with Environmental Resources Laboratory, Department of Landscape Architecture, Chungchou Institute of Technology, Changwa, 51003 Taiwan (ccchen@dragon.ccut.edu.tw)

Chiu-Yue Lin is with Department of Environmental Engineering and Science, Feng Chia University, Taichung, 40724, Taiwan (cylin@fcu.edu.tw).

Jaakko A. Puhakka is with Department of Chemistry and Bioengineering, Tampere University of Technology, Finland (jaakko.puhakka@tut.fi).

Olli Yli-Harja is with Department of Signal Processing, Tampere University of Technology, Finland (olli.yli-harja@tut.fi).

.

stoichiometric models [20]-[23] and models based on Anaerobic Digestion Model 1 (ADM1) [24]-[27] have been successfully used to describe the anaerobic bioprocesses. These models, however, require detailed *a priori* knowledge of the bioprocess [13], [23]. In bioprocesses involving mixed cultures, it is not always possible to establish detailed *a priori* knowledge about the bioprocess. It is very challenging to achieve the species specific growth and death rates in a mixed microbial community.

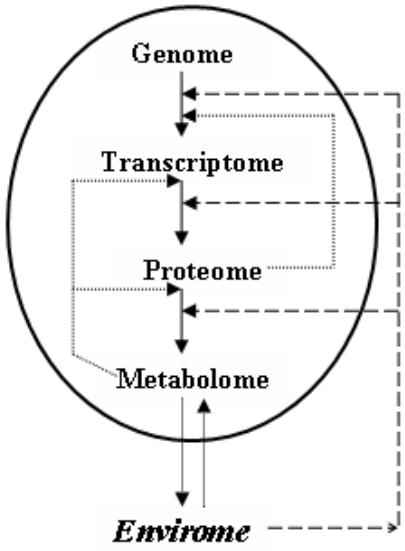


Fig. 1 Schematic representation of interactions between envirome and other omes. Adapted from [15].

In our earlier studies [28], [29], we have developed a clustering hybrid regression (CHR) approach which does not require detailed *a priori* knowledge of the bioprocess. The CHR approach offers means to reveal hidden patterns in bioprocess data; and to provide information for the optimization of $H_2$ production bioprocesses. The CHR approach has been successfully applied to xylose [28] and glucose [29] based fermentative $H_2$-producing bioprocesses. The clustering techniques (Self-Organizing Maps (SOM) and K-means) were applied to mine and cluster the end-products of fermentative $H_2$ production. The piecewise multiple linear regressions were used to model the $H_2$-production for each cluster.

In this study, we have extended the CHR approach by adding the prediction capabilities to it. The classification step (not included in CHR approach) has been added to provide the prediction capabilities. The developed predictive clustering hybrid regression (pCHR) approach models and predicts the $H_2$-production rate. The applicability of pCHR model was evaluated using dataset from $H_2$-producing sucrose-based bioreactor.

## II. Computational Methods

### A. Nonlinear Projection Pursuit

Projection pursuit (PP) methods [30]–[33] are computational techniques for extracting statistically significant features from high-dimensional datasets. PP techniques automatically determine the low-dimensional projections of such datasets and highlight any inherent clusters. Self-organizing maps and Sammon maps represent one subset (nonlinear projection pursuit) of more general PP techniques.

#### 1) Self Organizing Maps (SOM)

The SOM (also known as Kohonen's map) [34] is a data exploration tool for the analysis and visualization of nonlinear, high-dimensional data, and is generally used in the data understanding phase of the model development. It is a neural network algorithm based on unsupervised learning in a data-driven way, and combines the tasks of vector quantization and data projection [34], [35]. SOM has proven to be a valuable tool in data mining and knowledge discovery. It has been successfully applied in various engineering applications like image analysis, pattern recognition, chemical process monitoring and control, and fault diagnosis [34]–[38]. The capabilities of SOM in finding biologically meaningful clusters have also been demonstrated. The SOM has been used in clustering gene expression patterns from yeast or *C. elegans* [39]-[41] and also in cancer dataset [42]. The SOM has also found applications in nonlinear system identification and control [43]. In this study, SOM was used to reveal and visualize the relevant metabolic patterns and clusters in the bioprocess dataset.

SOM transforms an incoming signal pattern of arbitrary dimension into a one- or two- dimensional discrete map. The elements of the SOM are called neurons (also referred as map units or cells). Each neuron contains a reference vector (codebook vector), which is of the same dimension as the number of features (variables) in bioprocess data. The result of SOM is a topographic map of the input patterns in which the spatial locations (i.e., coordinates) of the neurons in the lattice are indicative of intrinsic statistical features contained in the input bioprocess dataset [34], [35].

SOM is optimized based on the quantization error (defined as equation 1) in the reference vector space. Let $\Phi : X \to A$ denote the SOM mapping from an input space $X$ to the discrete reference vector space $A$. Further, let $x \in X \in \Re^{n*1}$ be a bioprocess data sample, where n is the dimension of the bioprocess dataset. Let $m_i \in A \in \Re^{n*1}$ be $i$:th reference vector. The input bioprocess data sample ($x$) is connected to all neurons ($m_i$) in SOM, and the distances $d(x, m_i)$ between input bioprocess data sample and reference vectors are computed. The neuron having the closest reference vector ($m_c$) to the current input pattern is declared as winner (also referred as best matching unit) according to equation 1:

$$d(x, m_c) = \min_i (d(x, m_i)) \quad (1)$$

The winner is normally surrounded by a (topological) neighborhood region, $N_c$, and all neurons belonging to the neighborhood are updated. The updating is controlled by neighborhood function, $h_{ci}$, which is centered to the neuron having closest reference vector $m_c$. Let the number of reference vectors be $L$, then a distortion measure is defined as equation 2.

$$e(x) = \sum_{i=1}^{L} h_{ci} \cdot d(x, m_i) \quad (2)$$

In this study, free open source SOMPAK toolbox [44], [45] for MATLAB was used to plot SOM. The *a priori* parameters for SOM were set for default values as presented in [45].

*2) Sammon Maps*

Sammon's mapping (also referred to as non-linear mapping or NLM) is an iterative method based on a gradient search [46]. It is a non-linear mapping that maps a set of input points from a high-dimensional vector space onto a low-dimensional output space. The method attempts to preserve the inherent structure of the data when the patterns are projected from a high-dimensional space to low-dimensional space. The Sammon mapping is determined by the optimization of an error, or 'STRESS', measure which attempts to preserve all inter-point distances under the projection [47], [48]. The Sammon STRESS is defined as equation 3:

$$E = \frac{1}{\sum_{\mu=1}^{n-1} \sum_{v=\mu+1}^{n} d^*(\mu, v)} * \sum_{\mu=1}^{n-1} \sum_{v=\mu+1}^{n} \frac{\left[d^*(\mu, v) - d(\mu, v)\right]^2}{d^*(\mu, v)} \quad (3)$$

where $n$ is the number of patterns. The inter-pattern distances between pattern $\mu$ and pattern $v$ in the input space is $d^*(\mu, v)$ and in the projected output space is $d(\mu, v)$, respectively. These distance measures are generally Euclidean but need not be strictly so.

Sammon's mapping can be applied directly to multivariate data sets, but is computationally very intensive. It is thus applied when the SOM algorithm has already achieved a substantial data reduction by replacing the original data vectors with a smaller number of representative prototype codebook vectors. The resulting Sammon visualization depicts clusters in input space as groups of data points mapped close to each other in the output space. Thus, the inherent structure of the input patterns can be told from the structure detected in the 2-dimensional visualization.

In this study, Sammon function implementation in free open source SOMPAK toolbox [44], [45] for MATLAB was used to plot Sammon map. The *a priori* parameters for Sammon were set for default values as presented in [45].

*B. K-means clustering*

K-means [49] clustering is a simple unsupervised learning algorithm used to solve clustering problems. It classifies a given dataset into a certain number of clusters ($k$) which are fixed *a priori* [50], [51]. In this research, SOM and Sammon visualizations were used to guide in choosing this $k$ value.

Steps of K-means clustering include: (i) Choosing the number of clusters ($k$), (ii) randomly choosing the $k$ cluster centers (known as centroids), (iii) measuring the distance of objects to centroids and grouping them based on minimal distances, (iv) if any objects moves the group, go back to step (ii). K-means is a simple algorithm that has been adapted to many problem domains. The algorithm is significantly sensitive to the initial randomly selected centroids. Thus, the K-means algorithm should be run multiple times to reduce this effect.

Several distance metrics exist for calculating the distances of objects to centroids [50], [51]. It is a difficult decision to choose the distance measure when using a clustering algorithm. The standard choice is the Euclidean distance, as its simple. There are many situations in bioprocess data analysis where Euclidean distance may not be the best choice. The reason being, that from a biotechnological point of view, the direction of the change in metabolic profile is very often more important than the difference between the ratios [29]. Euclidean distance is incapable to take direction into account. The correlation distance measure is capable of taking also directions of the changes in profile into account.

To cluster the dataset, MATLAB *v. 7.3* function *kmeans*, with correlation as distance measure was used.

*C. K-Nearest Neighbor (kNN) Classifier*

K-nearest neighbor (kNN) algorithm [52], a variation of nearest-neighbor, is an instance-based nonparametric classification technique with successful applications in statistical pattern recognition problem. The kNN is well suited for multi-modal classes (i.e., consists of objects whose independent variables have different characteristics for different subsets) as its classification decision is based on a small neighborhood of similar objects [53], [54].

Let $y \in \{y_1, y_2, y_3, ...y_m\}$ be an unlabelled input pattern to be classified to one of the class $c \in \{c_1, c_2, c_3, ...c_i\}$. Let $x \in \{x_1, x_2, x_3, ...x_n\}$ be a pattern (already clustered dataset in our case) which have label $c$. In nearest neighbor only the point $x$ that is closest to $y$ is computed and the class $c_i$ of $x$ is the class of $y$. In kNN, $k$ points that are closest to $y$ are calculated. The unlabelled input pattern $y$ is assigned to the class $c_i$, if $c_i$ has the biggest similarity score (or majority score) to $y$ among all classes. Equations 4 and 5 are the widely used strategies for kNN classification.

$$l(y_i) = \arg\max_k \sum_{x_j \in kNN} l(x_j, c_k) \quad (4)$$

$$l(y_i) = \arg\max_k \sum_{x_j \in kNN} sim(y_i, x_j) l(x_j, c_k) \quad (5)$$

where $y_i$ is an unlabelled input pattern, $x_j$ is one of the computed $k$ nearest neighbors, $l(x_j, c_k) \in \{0,1\}$ indicates whether $x_j$ belongs to class $c_k$ and $sim(y_i, x_j)$ is the similarity function for $y_i$ and $x_j$. Equation (4) means that the winner class will be the class that has the largest number of members in the $k$ nearest neighbors; whereas equation (5) means the class with maximal sum of similarity will be the winner.

To classify the dataset, MATLAB *v. 7.3* function *knnclassify*, with correlation as distance measure was used.

*D. Silhouette Plots*

Silhouette plots were used to display and evaluate the clustering and classification results. Consider a node $v_i$ that belongs to cluster $C_j$. Let $C_h$ be the closest (according to average distance) cluster to node $v_i$. The silhouette index [55], [56] is defined as equation 6.

$$s(v_i) = \frac{d(v_i, C_h) - d(v_i, C_j)}{\max\left(d(v_i, C_j), d(v_i, C_h)\right)} \quad (6)$$

and $-1 \le s(v_i) \le 1$

where $d(v_i, C_j)$ is the average dissimilarity of $v_i$ object to all other objects in the same cluster; and $d(v_i, C_h)$ is the minimum of average dissimilarity of $v_i$ object to all objects in other cluster (in the closest cluster).

When $s(v_i)$ is close to 1, $v_i$ is said to be "well clustered". When $s(v_i)$ is close to 0, $v_i$ is said to be intermediate between two clusters. When $s(v_i)$ is close to -1, $v_i$ is said to be "badly clustered". The largest overall average silhouette indicates the best clustering (the number of clusters).

The clusters were validated using *silhouette* function in MATLAB *v. 7.3*.

*E. Predictive Clustering Hybrid Regression (pCHR)*

The proposed pCHR model is an extension to the CHR approach [28], [29]. In pCHR approach, the prediction capabilities have been introduced by adding classification step to the CHR approach. The pCHR approach combines the ideas from clustering [50], [51], classification [52]-[54] and regression techniques [57]–[60].

*1) General schema*

Fig. 2 presents the general schema and idea of pCHR approach. The dataset is divided into two parts for training and testing (prediction) purposes. Training dataset is sub-grouped into clusters based on their statistical (or functional, if known) features. The modeling approach is chosen and applied to the clusters obtained. Based on the clusters, testing dataset is then classified (class label is predicted). The parameters of the model chosen in training phase are applied to classified dataset and criterion variable is predicted using predictor variables.

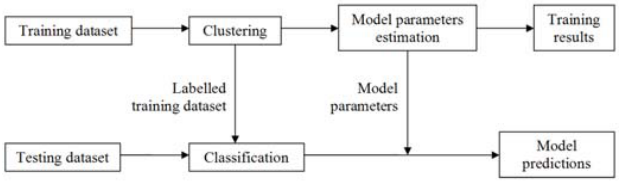


Fig. 2 General schema for pCHR approach.

The idea of pCHR model doesn't restrict to computational techniques applied in this study. The neural networks based techniques (adaptive resonance theory and multi layer perceptrons) for clustering/classification of glucose based $H_2$-dataset have been evaluated [61]. The models based on SOM clustering have also been evaluated for $H_2$-dataset [29]. For $H_2$ production datasets [28], [29], K-means clustering gives optimal and stable clustering results. K-means clustering is used in this study, to keep the computational complexity low. Several other clustering and classification techniques [50]-[54] can be applied and need to be evaluated. The modeling paradigm (here piecewise multiple linear regression) is also an open choice, depending on the properties of the dataset available, and the objective of modeling [16], [62].

*2) Algorithm of pCHR*

Fig. 3 presents the flow chart of pCHR algorithm as applied in this study. The *k*-means clustering was used to cluster the training dataset, while kNN classifier was used to classify test dataset. Multiple piecewise linear regression was used to obtain local regression models for each cluster.

Piecewise linear regression is a local modeling approach that proposes different straight-line relationships for different intervals over the range of data [57]. Breakpoints which define the interval boundaries are the values where the slope of the linear function changes. The regression function at the breakpoint maybe discontinuous, but a model can be written in such a way that the function is continuous at all points including the breakpoints [59].

In pCHR approach, clusters obtained (for training dataset) from K-means clustering define these intervals (subsets of data). The relationships between the response and the explanatory variables are then modeled. The model has different regression parameter values for different clusters. For each of the clusters obtained, multiple regressions are done to analyze the relationship between variables. The computational problem that needs to be solved in multiple regression analysis is to fit a straight line (or plane in an *n*-dimensional space, where *n* is the number of independent variables) to a number of points [58], [60]. The mathematical form of pCHR model is given as equation 7.

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_k \end{bmatrix} = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1n} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ \beta_{k1} & \beta_{k2} & \cdots & \beta_{kn} \end{bmatrix} \times \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_k \end{bmatrix} \quad (7)$$

where $k$ is the number of clusters and $n$ is the number of predictor variables. $\beta$ is the parameters of the model, $X$ is the predictor variable and $\varepsilon$ is the noise term.

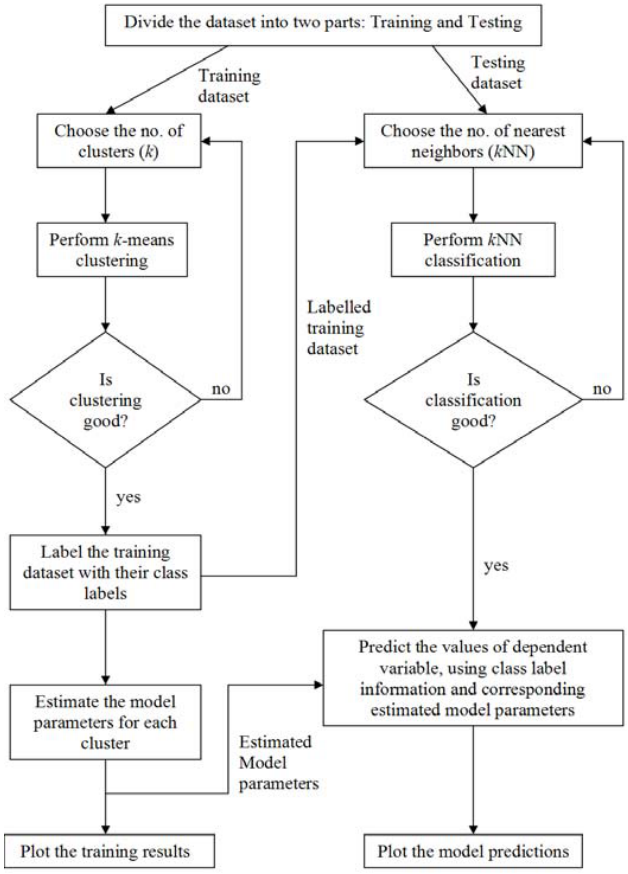


Fig. 3 Algorithm of pCHR model as used in this study.

To evaluate the prediction capabilities, unseen testing dataset was used. The unseen dataset is presented to the model. The kNN classifier is used to determine the class labels of the incoming samples. The parameters ($\beta$ and $\varepsilon$ obtained earlier) corresponding to the class label are used to predict the criterion variable.

In this study, applicability of pCHR approach to predict sucrose-based $H_2$ production (experimental setup described in next section) was evaluated. $H_2$ production rate was modeled as a function of *metabolome* ($CO_2$-production rate, concentrations of acids and alcohols) and *envirome* (pH, hydraulic retention time (HRT), oxidation-reduction potential (ORP), alkalinity, recycle ratio, sucrose concentration, sucrose degradation and biomass). The bioprocess dataset consisted of 278 sample points. Table I describes the composition and dynamic ranges of the dataset.

TABLE I COMPOSITION, DYNAMIC RANGES AND MEASUREMENT UNITS OF THE BIOPROCESS DATA USED IN THE STUDY

| Variables | Min | Mean | Max |
|---|---|---|---|
| **Recycle ratio (%)** | 0 | 29.24 | 100 |
| **HRT (h)** | 2 | 10.20 | 12 |
| **sucrose conc. (g COD/l)** | 20 | 23.79 | 40 |
| **sucrose degr. (%)** | 44.39 | 87.62 | 100 |
| **Biomass (g VSS/l)** | 1.33 | 3.89 | 6.81 |
| **pH** | 5.59 | 6.64 | 7.19 |
| **ALK (mg/L as CaCO3)** | 2680 | 4810.16 | 7290 |
| **ORP (mV)** | -613 | -444.55 | -216 |
| **EtOH (mg COD/L)** | 136.13 | 2810.60 | 8796.14 |
| **HAc (mg COD/L)** | 183.78 | 2335.62 | 4583.09 |
| **HPr (mg COD/L)** | 0 | 739.43 | 3520.29 |
| **HBu (mg COD/L)** | 446.79 | 5334.11 | 14000 |
| **$H_2$PR (l/h/l)** | 0 | 0.22 | 0.92 |
| **$CO_2$PR (l/h/l)** | 0 | 0.30 | 0.88 |

HRT: hydraulic retention time, conc.: concentration, degr.: degradation, ALK: alkalinity, ORP: oxidation-reduction potential, EtOH: ethanol, HAc: acetate, HPr: propionate, HBu: butyrate, $H_2$PR: hydrogen production rate, $CO_2$PR: carbon dioxide production rate, Min: minimum, Max: maximum.

The function *regstats* from MATLAB was used to estimate the parameters ($\beta$ and $\varepsilon$) of the model. The parameter values obtained for pCHR model are shown in Table II.

## III. Case Study: Sucrose-based Biohydrogen Production

To evaluate the applicability of pCHR approach, sucrose-based fermentative biohydrogen dataset was used as a case study. The dataset was obtained by operating a laboratory-scale experimental system as shown in Fig. 4. The details of the seed sludge, substrate, reactor setup, assessing protocol and analyses methods are presented below.

### A. Seed Sludge

The Li-Ming Municipal Sewage Treatment Plant (Taichung, Taiwan) supplied the seed sludge for this study. The seed sludge was collected from the final sedimentation tank. The collected sludge was screened with a No. 8 mesh (diam. 2.35 mm) and was heated at 100°C for 45 minutes to inhibit methanogen or other microorganisms' bioactivity.

### B. Substrate

The seed sludge was acclimated with sucrose at a concentration of 20000 mg COD/L. The substrate contained sufficient inorganic [63] (mg/L): $NH_4HCO_3$ 5240, $K_2HPO_4$ 125, $MgCl_2{\cdot}6H_2O$ 100, $MnSO_4{\cdot}6H_2O$ 15, $FeSO_4{\cdot}7H_2O$ 25,

$CuSO_4 \cdot 5H_2O$ 5, $CoCl_2 \cdot 5H_2O$ 0.125, $NaHCO_3$ 6720. The substrate was stored at 4°C.

TABLE II PCHR MODEL PARAMETERS VALUES OBTAINED USING THE MATLAB *REGSTATS* FUNCTION

| parameters | cluster 1 | cluster 2 | cluster 3 |
|---|---|---|---|
| $\varepsilon$ | 1.2737 | -0.0954 | 0.1923 |
| $\beta_{Recycle\ ratio}$ | -6.3685 | 0.0002 | 0.0008 |
| $\beta_{HRT}$ | -0.0872 | -0.0201 | 0.0026 |
| $\beta_{sucrose\ conc.}$ | 0.0105 | 0.0043 | -0.0016 |
| $\beta_{sucrose\ degr.}$ | 0.0045 | 0.0081 | -0.0004 |
| $\beta_{Biomass}$ | 0.043 | -0.0205 | -0.0028 |
| $\beta_{pH}$ | 0.6193 | -0.1055 | -0.0242 |
| $\beta_{ALK}$ | -0.0001 | 0 | 0 |
| $\beta_{ORP}$ | -0.0008 | 0 | 0.0001 |
| $\beta_{EtOH}$ | 0.0001 | 0 | 0 |
| $\beta_{HAc}$ | -0.0002 | 0 | 0 |
| $\beta_{HPr}$ | 0.0001 | 0 | -0.0001 |
| $\beta_{HBu}$ | 0.0001 | 0 | 0 |
| $\beta_{CO2PR}$ | -0.0491 | 0.6894 | 0.938 |

HRT: hydraulic retention time, conc.: concentration, degr.: degradation, ALK: alkalinity, ORP: oxidation-reduction potential, EtOH: ethanol, HAc: acetate, HPr: propionate, HBu: butyrate, $CO_2$PR: carbon dioxide production rate

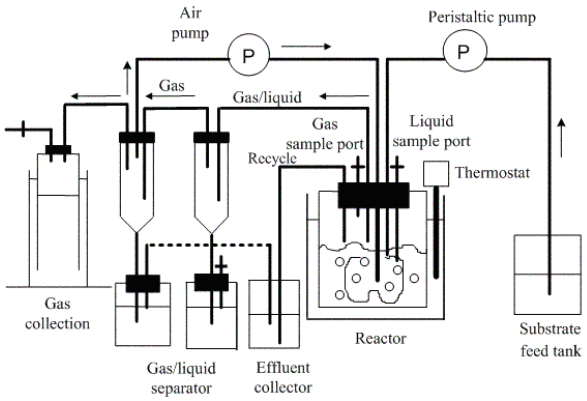


Fig. 4 Configuration of an anaerobic bioreactor for continuous $H_2$ production.

### C. Completely Stirred Tank Reactor (CSTR)

The CSTR (Fig. 4), with a working volume of 4.0 L, was operated for 15 months. The reactors were placed in a water-bath tank and maintained the operating temperature 35±1 °C. The substrate was drawn out with a peristaltic pump from the refrigerator through the 35 °C water-bath tank then into CSTR. The overflow and digestion biogas were drawn out with a diaphragm motor (air pump) from reactor. When the overflow and digestion biogas passed through the gas/liquid separator, the overflow was collected and the biogas was circulated in the reactor. Each reactor was connected to a gas collection cylinder placed in a saturated salt solution. Each cylinder is same working volume (300 mL) and equipped with a counter that used to calculate the total volume of produced biogas. The effluent was collected in an Effluent Collector from which the settled sludge was recycled into the reactor to maintain biomass concentrations. The effluent recycle ratio was defined as recycle velocity dived by the influent velocity [64].

### D. Assessing Protocol

Initial CSTR operation was in a continuous feeding mode and hydraulic retention time (HRT) was 12 h. pH was controlled around 6.7 which was found to be favorable for hydrogen production [65], [66]. When a steady-state condition was reached and the desired data were obtained the recycle ratio or substrate concentration or HRT was reduced. At each run, the CSTR was operated for more than ten times of the HRT to develop a steady-state condition. Steady-state conditions reached when the product concentrations such hydrogen gas content, biogas volume and metabolite concentrations were stable (less than 10 % variation). For each steady-state data measurement, 6-10 samples were determined.

During the experimental operation of this study, the reactor was routinely monitored for pH, alkalinity, oxidation-reduction potential (ORP), gas production and composition, sucrose concentration, ethanol concentration, volatile fatty acid (VFA) distribution and VSS concentrations. The gas volumes were corrected to a standard temperature (0 °C) and pressure (760mmHg) (STP).

### E. Bioprocess Monitoring Analyses

The mixed liquors sampled were centrifuged (900 g, 15 min) and the supernatants were taken for metabolite analysis. VFA and ethanol were analyzed with a gas chromatograph having a flame ionization detector (Shimadze GC-14, Japan). Biogas volume was determined by a gas meter (Ritter, Germany). Biogas composition except hydrogen sulfide was analyzed with a gas chromatograph having and a thermal conductivity detector (China Chromatograph 8700T, Taiwan). Hydrogen sulfide gas was analyzed with a gas chromatograph having a flame photometric detector (capillary column, 150 °C; injection temperature, 150 °C; carrier gas, $N_2$). Other analytical details for the VFA, ethanol and biogas assays were the same as those in [67], [68]. Anthrone-sulfuric acid method was used to measure sucrose [69]. The ORP value was measured using a pH/ORP Controller with a silver chloride electrode (Suntex, Taiwan). Other water quality parameters were measured according the procedures of Standard Methods [70].

## IV. Results and Discussion

### A. Bioprocess data analyses

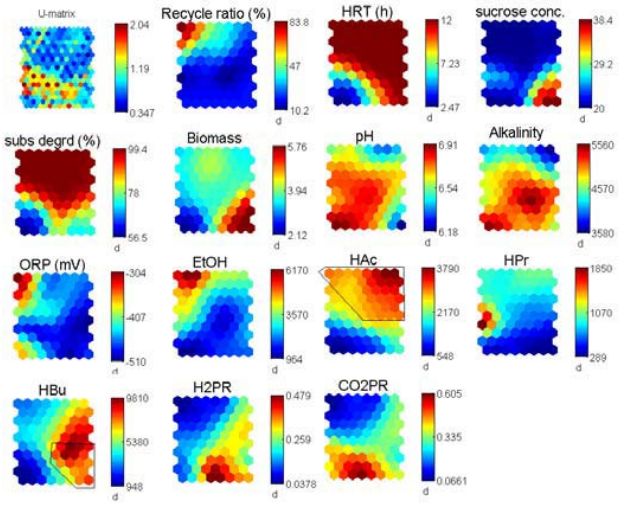


Fig. 5 SOM visualization (with colored component planes) of bioprocess dataset. Dark red represents the highest value observed for the variable. Dark blue represents the lowest value observed for the variable. Two metabolic patterns of acetate coupled with other metabolites, and butyrate alone are highlighted. Remaining neurons represent the transition state.

SOM maps with colored component planes (Fig. 5) were used to visualize and identify the metabolic patterns and clusters of the fermentation process during the CSTR operation. Colored component planes were formed from the SOM reference vector by splitting it to n components, where n (for our dataset n = 14) is dimension of the reference vector. In context of bioprocess data, component plane corresponds to a feature (variable) in the dataset. Neurons in the component planes were color shaded. The shades of red correspond to high values for the variable, the shades of yellow to moderate values and the shades of blue to low values.

Hydrogen production is accompanied with VFAs or solvent production during a dark fermentation. These liquid metabolites change in concentration distributions and fractions when the operation conditions such as cultivation pH, temperature, HRT and substrate concentration are changed [6], [8], [14]. Even when the operational conditions are fixed, the change in the microbial community structure also changes the metabolic profile and distributions [71], [72]. Liquid product analysis shows that the major VFAs were acetate, propionate and butyrate, with butyrate as the major component. Moreover, the liquid metabolite concentrations varied when HRT and recycle ratio were changed.

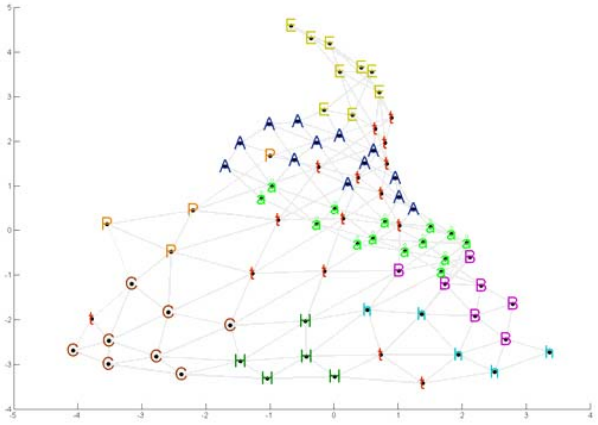

Fig. 6 Sammon map visualization of nine metabolic profiles observed. t: transition phase, C: carbon dioxide, P: propionate, E: ethanol coupled with acetate, A: acetate alone, å: butyrate coupled with acetate, B: butyrate alone, h: hydrogen coupled with butyrate, and H: hydrogen alone.

SOM analyses suggested three distinct metabolic patterns (clusters): 1) acetate coupled with other metabolites (ethanol and butyrate), 2) butyrate alone, and 3) transition state when no metabolite was dominating. High $H_2$-production rates are usually seen with high butyrate production. In our case, $H_2$ production rates were moderate during high butyrate phase. This maybe likely due to $H_2$ consuming species present during this metabolic phase. Acetate, propionate and ethanol are not high $H_2$ producing reactions [6], [8], [14], and it can also be seen from SOM analyses. In our case study, best $H_2$ production rates (0.92 l/h/l) were observed when the HRT was 6 hrs, alkalinity was 4870 (mg/l as CaCO3), ORP was -409 mV, pH was around 6.7 and with lower concentrations of sucrose (20 gCOD/l). High concentrations of substrate may inhibit the metabolism [13]. Also substrate degradation of around 76% and low biomass (3.5 gVSS/l) was observed during high $H_2$ production phase. Theoretically, high substrate degradation and high biomass is associated with high $H_2$ production [6], [8], [14], [72]. High biomass (in mixed microbial community) not always signifies high $H_2$ production. It also depends on the microbial structure of the community [71]. The dynamics of the microbial community structure during the bioprocess operation needs to be analyzed to understand the relationships between several metabolites.

SOM clusters were further analyzed to obtain insight into the three dominant metabolic clusters. The closer analyses of the metabolic clusters suggested nine metabolic profiles. These metabolic profiles were, however, not considered to be separate clusters, as they were found to be topologically close to each other in SOM and Sammon map. The profiles observed were: 1) transition phase (t), 2) carbon dioxide, (C) 3) propionate (P), 4) ethanol coupled with acetate (E), 5) acetate alone (A), 6) butyrate coupled with acetate (å), 7) butyrate alone (B), 8) hydrogen coupled with butyrate (h), and 9) hydrogen alone (H). The profiles are in the sequence as presented in Fig. 8, where the numbers of nine metabolic

patterns (marked on Y-axis of Fig. 8) corresponds to these profiles. The symbolic code in parentheses is used for representation in the Sammon map (Fig. 6), used to visualize these metabolic profiles.

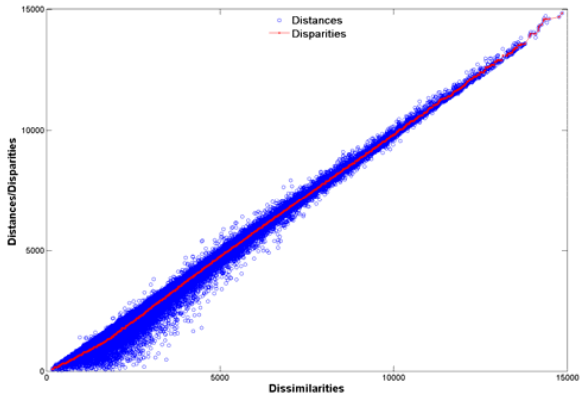


Fig. 7 Shepard plot (with Sammon stress as scaling criterion) of the dataset. The distances in scaling approximate the disparities (the scatter of blue circles about the red line), and the disparities reflect the ranks of the dissimilarities (the red line is linear but increasing and becomes slightly nonlinear at the tail).

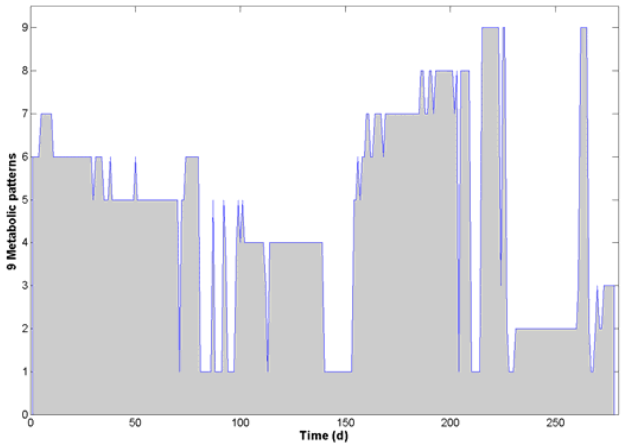


Fig. 8 Time series visualization of metabolic profiles and their transitions in bioprocess data. The nine metabolic profiles observed are marked in following sequence on Y-axis: 1: transition phase, 2: carbon dioxide, 3: propionate, 4: ethanol coupled with acetate, 5: acetate alone, 6: butyrate coupled with acetate, 7: butyrate alone, 8: hydrogen coupled with butyrate, and 9: hydrogen alone.

The Shepard plot (with Sammon stress as scaling criterion) was used to visualize the distances/disparities and the ranks of the dissimilarities in the dataset. The Shepard plot is a scatter-plot of the interpoint distances vs. the original dissimilarities, and is used to determine the goodness of fit. Fig. 7 shows a linear pattern and a small scatter of data points. It implies that the distance in Sammon visualization is a good reflection of dissimilarity in our dataset.

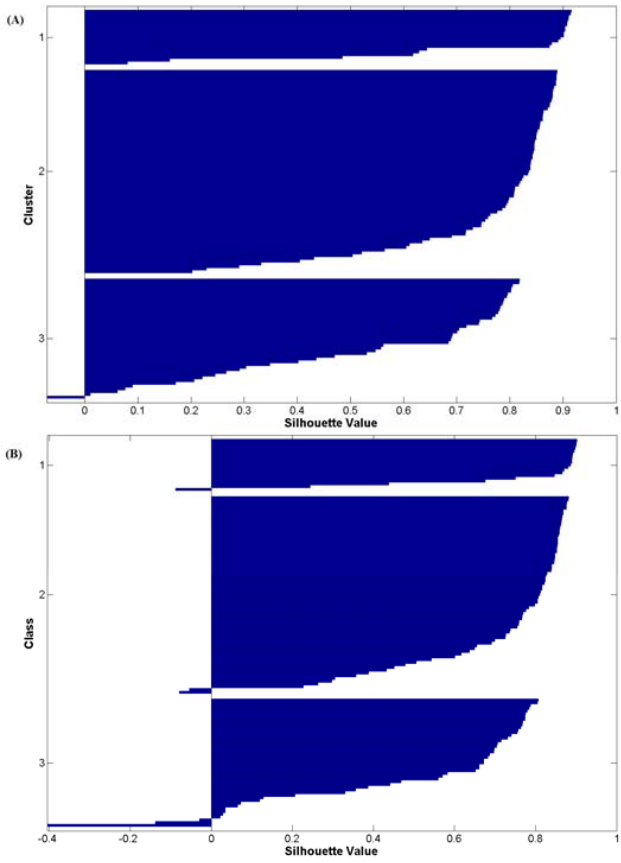


Fig. 9 Silhouette plots (A) K-means clustering over training dataset (mean silhouette value 0.69), (B) kNN classification over test dataset (mean silhouette value 0.65)

Time series of the bioprocess dataset was analyzed. In this study, we were interested in understanding the dominant metabolic state changes. Instead of plotting each metabolite separately, the dominant metabolic profiles as observed were plotted as time series. SOM best matching units were used to label each input pattern to one of the 9 observed profiles. These metabolic profiles are plotted as time series in Fig. 8. The numbers on the Y-axis of Fig. 8 represents the observed metabolic profiles. As shown in the Fig. 8, the bioprocess started with butyrate (marked as 7 on Y-axis) and butyrate coupled with acetate (marked as 6 on Y-axis). Later it moved to acetate dominating metabolic phase (marked as 5 on Y-axis). After few transition states (marked as 1 on Y-axis), the bioprocess shifted to ethanol coupled with acetate state (marked as 4 on Y-axis). The transition again occurred and bioprocess moved to butyrate dominating state (marked as 7 on Y-axis) and butyrate coupled with hydrogen state (marked as 8 on Y-axis). At this phase, the high $H_2$-production state (marked as 9 on Y-axis) was observed. Finally the bioprocess moved to carbon dioxide producing state (marked as 2 on Y-axis). The bioprocess operation was terminated when propionate (marked as 3 on Y-axis) started to appear in high amounts. The study of the microbial community dynamics is required to better understand the metabolic transitions during

bioreactor operation [71].

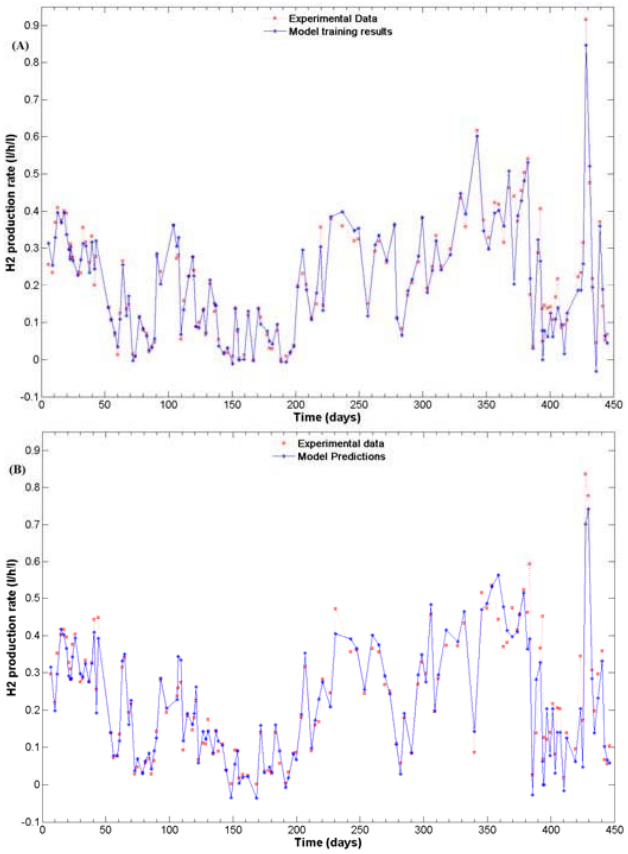


Fig. 10 pCHR model results **(A)** training dataset (MSE = 0.0014), **(B)** model predictions (MSE = 0.0032).

### *B. pCHR model*

The bioprocess dataset (consist of 278 samples) was divided into two parts for training (139 samples) and testing (139 samples) purposes. The clusters in training dataset were obtained using K-means clustering algorithm with correlation as distance measure (Fig. 9a). SOM and Sammon analyses suggested three distinct metabolic clusters, thus *k=3* were chosen for K-means clustering. The clusters were evaluated using Silhouette index, and the mean silhouette value obtained was 0.69 when *k* was chosen to be 3. Increasing or decreasing the value of *k* resulted in poor mean silhouette values. The clusters thus obtained were used to classify the data patterns in testing dataset. The classification of test dataset (Fig. 9b) was done using kNN classifier (with correlation as distance measure and number of nearest neighbors as 6) and the mean silhouette value obtained was 0.65.

The parameters for pCHR model were obtained by applying multiple regressions over clustered training dataset. The parameters, thus obtained, were used for predicting $H_2$ production rates. The results show that pCHR model (Figs. 10a, b) performed very well in predicting the $H_2$-production rate. The mean square error (MSE) for training was 0.0014 and 0.0032 for testing

## V. Conclusions

A predictive clustering hybrid regression (pCHR) approach was used to model and predict $H_2$-production rate based on the metabolic data, control parameters and envirome variables. SOM component planes and Sammon map were used to visualize clusters and to study correlations between bioprocess data samples (envirome and metabolome). SOM and Sammon analyses detected nine distinct metabolic profiles (acetate, acetate coupled with ethanol, acetate coupled with butyrate, butyrate, butyrate coupled with hydrogen, hydrogen, carbon dioxide, propionate and transition state). Three clusters (acetate coupled with other metabolites, butyrate and transition phase) were observed in the bioprocess dataset. The optimal envirome conditions obtained for this bioprocess were: HRT of 6 hrs, alkalinity as 4870 (mg/l as CaCO3), ORP as -409 mV, pH as 6.7 and concentrations of sucrose as 20 gCOD/l. The pCHR model performed very well (MSE of 0.0014 and 0.0032 for training and testing, respectively) in modeling the $H_2$-production rate.

## Acknowledgment

The authors acknowledge the support by HydrogenE (research project (2005 - 2008) under Academy of Finland, application number 107425). The work was also supported by the Academy of Finland (application number 213462, Finnish Programme for Centers of Excellence in Research 2006-2011) and National Science Council of Taiwan, R.O.C. (Contract No. NSC 91-2211-E-235-002). Authors would like to thank Ms. Chia-Jung Tsai and Mr. Chyi-How Lay from Feng Chia University, Taiwan for providing the experimental dataset.

## References

[1] H. Kitano, *Foundations of Systems Biology*. The MIT Press, 2001.
[2] M. T. Facciotti, R. Bonneau, L. Hood, and N. S. Baliga, "Systems biology experimental design - considerations for building predictive gene regulatory network models for prokaryotic systems," *Current Genomics*, vol. 5, no. 7, pp. 527–544, Nov. 2004.
[3] H. Kitano, "Systems biology: a brief overview," *Science*, vol. 295, no. 5560, pp. 1662–1664, March 2002.
[4] A. Kremling, and J. Saez-Rodriguez, "Systems biology – an engineering perspective," *J. Biotechnol.*, vol. 129, pp. 329–351, 2007.
[5] R. Takors, B. Bathe, M. Rieping, S. Hans, R. Kelle, and K. Hutchmacher, "Systems biology for industrial strains and fermentation processes – example: amino acids," *J. Biotechnol.*, vol. 129, pp. 181–190, 2007.
[6] P.C. Hallenbeck, "Fundamentals of fermentative production of hydrogen," *Water Sci. Technol.,* vol. 52, no. 1-2, pp. 21–29, 2005.
[7] J'O. M.Bockris, "The origin of ideas on a hydrogen economy and its solution to the decay of the environment," *Int. J. Hydrogen Energy*, vol. 27, pp. 731–740, 2002.
[8] D. Das, and T.N. Veziroğlu, "Hydrogen production by biological processes: a survey of literature," *Int. J. Hydrogen Energy*, vol. 26, pp. 13–28, 2001.
[9] J. Benemann, "Hydrogen biotechnology: progress and prospects," *Nat. Biotechnol.*, vol. 14, pp. 1101–1103, 1996.
[10] I. K. Kapdan, and F. Kargi, "Bio-hydrogen production from waste materials," *Enzyme Microb. Tech.*, vol. 38, pp. 569–582, 2006.
[11] C. Li, and H. H. P. Fang, "Fermentative hydrogen production and wastewater and solid wastes by mixed cultures," *Crit. Rew. Env. Sci. Technol.*, vol. 37, pp. 1–39, 2007.

[12] C.-Y. Lin, and R.-C. Chang, "Fermentative hydrogen production at ambient temperature," *Int. J. Hydrogen Energy,* vol. 29, pp. 715–720, 2004.
[13] J. Rodriguez, R. Kleerebezem, J. M. Lema, and M. C. van Loosdrecht, "Modeling product formation in anaerobic mixed culture fermentations," *Biotechnol. Bioeng*., vol. 93, pp. 592–606, 2006.
[14] R. Nandi, and S. Sengupta, "Microbial production of hydrogen: an overview," *Crit. Rev. Microbiol.,* vol. 24, pp. 61–84, 1998.
[15] G. Liden, "Understanding the bioreactor," *Bioprocess Biosyst. Eng.*, vol. 24, pp. 273– 279, 2002.
[16] Nikhil, "Formulation of mathematical models for control and optimization of bioreactors," M.Sc. thesis, Dept. Environmental Technology, Tampere Univ. Technology, Tampere, Finland, 2005.
[17] Nikhil, "Application of systems bioengineering for fermentative hydrogen production," presented at 3rd TICSP Workshop on Computational Systems Biology, WCSB 2005, Tampere, Finland, June 13 – 14, 2005, pp. 33–34.
[18] K. Y. Rani, and V. S. R. Rao, "Control of fermenters – a review," *Bioprocess Eng.*, vol. 21, pp. 77–78, 1999.
[19] Schugerl, K.; Bellgardt, K.H. *Bioreaction engineering. Modeling and control.* Berlin, Heidelberg, New York: Springer-Verlag. 2000.
[20] Bailey, E.J. Mathematical modeling and analysis in biochemical engineering: Past accomplishments and future opportunities. *Biotechnol. Prog.* 1998, 14, 8-20.
[21] Bernard, O.; Bastin, G. On the estimation of the pseudo-stoichiometric matrix for macroscopic mass balance modelling of biotechnological processes. *Math. Biosci.* 2005, 193, 51-77.
[22] Husain, A. Mathematical models of the kinetics of anaerobic digestion – a selected review. *Biomass. Bioenerg.* 1998, 14, 561-571.
[23] McCarty, P.L.; Mosey, F.E. Modelling of anaerobic digestion processes (a discussion of concepts). *Wat. Sci. Technol.* 1991, 24:8, 123-129.
[24] Batstone, D.J.; Keller, J.; Angelidaki, I.; Kalyuzhnyi, S.V.; Pavlostathis, S.G.; Rozzi, A.; Sanders, W.T.M.; Siegrist, H.; Vavilin, V.A. *Anaerobic digestion model no. 1 (ADM1)*, IWA Task Group for mathematical modelling of anaerobic digestion processes. London, UK: IWA Publishing 2002.
[25] Blumensaat, F.; Keller J. Modelling of two-stage anaerobic digestion using the IWA Anaerobic Digestion Model No. 1 (ADM1). *Water Res* 2005, 39, 171-183.
[26] Kalyuzhnyi, S.V. Batch anaerobic digestion of glucose and its mathematical modeling. II. Description, verification and application of model. *Bioresour. Technol.* 1997, 59, 249-258.
[27] Parker, W.J. Application of the ADM1 model to advanced anaerobic digestion. *Bioresour. Technol.* 2005, 96, 832-1842.
[28] Nikhil, A. Visa, O. Yli-Harja, C.-Y. Lin, and J. A. Puhakka, "Application of the Clustering Hybrid Regression Approach to Model Xylose-Based Fermentative Hydrogen Production," *Energy Fuels*, 2008, 22 (1), 128–133.
[29] Nikhil, P. E. P. Koskinen, A. Visa, A. H. Kaksonen, J. A. Puhakka, and O. Yli-Harja, "Clustering hybrid regression (CHR): a novel computational approach to study and model biohydrogen production through dark fermentation," *Bioprocess and Biosystems Engineering,* 2008*,* doi: 10.1007/s00449-008-0213-9.
[30] P. J. Huber, "Projection pursuit," *Ann. Statist.*, vol. 13, no. 2, pp. 435–475, 1985.
[31] J. H. Friedman, "Exploratory projection pursuit," *J. Amer. Statist. Assoc.*, vol. 82, no. 397, pp. 249–266, 1987.
[32] B. D. Ripley, "Neural networks: a review from statistical perspective," *Statistical Sci.*, vol. 9, no. 1, pp. 45–48, Feb. 1994.
[33] J. A. Lee, A. Lendasse, and M. Verleysen, "Nonlinear projection with curvilinear distances: isomap versus curvilinear distance analysis," *Neurocomputing*, vol. 57, pp. 49–76, 2004.
[34] T. Kohonen, *Self-organizing maps.* Springer, Berlin, Heidelberg, New York: Springer Series in Information Sciences, vol. 30, 1995.
[35] S. Kaski, "Data exploration using self-organizing maps," D.Tech. (Ph.D.) dissertation, Helsinki University of Technology, Finland, 1997.
[36] M. Kasslin, J. Kangas, and O. Simula, "Process state monitoring using self organizing maps," in *Artificial Neural Networks*, vol. 2, I. Aleksander, and J. Taylor, Eds. Amsterdam, The Netherlands, North Holland, 1992, pp. 1531–1534.
[37] O. Simula, and J. Kangas, *Process monitoring and visualization using self-organizing maps. Neural networks for chemical engineers. Computer-aided chemical engineering*. Amsterdam: Elsevier, 1995, pp. 377–390.
[38] H. Yin, "ViSOM – a novel method for multivariate data projection and structure visualization," *IEEE Trans. Neural Networks*, vol. 13, no. 1, pp. 237–243, Jan. 2002.
[39] Tamayo, P.; Slonim, D.; Mesirov, J.; Zhu, Q.; Kitareewan, S.; Dmitrovsky, E.; Lander, E.; Golub, T. Interpreting patterns of gene expression with self-organizing maps; methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences*, USA 1999, 96, 2907-2912.
[40] Törönen, P.; Kolehmainen, M.; Wong, G.; Castren, E. Analysis of gene expression data using self-organizing maps. *FEBS Letters* 1999, 451:2, 142-146.
[41] Hill, A.; Hunter, C.; Tsung, B.; Tucker-Kellogg, G.; Brown, E. Genomic analysis of gene expression in C.elegans. *Science* 2000, 290, 809-812.
[42] Chen, D.-R.; Chang, R.-F.; Huang, Y.-L. Breast cancer diagnosis using self-organizing maps for sonography. *Ultrasound in Medicine and Biology* 2000, 26:3, 405-411.
[43] J. C. Principe, L. Wang, and M. A. Motter, "Local dynamic modeling with self-organizing maps and applications to nonlinear system identification and control," *Proc. IEEE*, vol. 86, no. 11, pp. 2240–2258, Nov. 1998.
[44] T. Kohonen, J. Hynninen, J. Kangas, and J. Laaksonen, "SOM_PAK: The self-organizing map program package," Laboratory of Computer and Information Science, Helsinki University of Technology, Finland, Technical Report A31, 1996.
[45] J. Vesanto, J. Himberg, E. Alhoniemi, and J. Parhankangas, (2000) "SOM Toolbox for MATLAB 5," SOM Toolbox Team, Helsinki University of Technology, Finland. Available: http://www.cis.hut.fi/projects/somtoolbox/.
[46] J. W. Sammon, Jr, "A nonlinear mapping for data structure analysis," *IEEE Trans. Computers*, vol. c-18, no. 5, pp. 401–409. May 1969.
[47] D. K. Agrafiotis, "A new method for analyzing protein sequence relationships based on Sammon maps," *Protein Sci.*, vol. 6, pp. 287–293, 1997.
[48] B. Lerner, H. Guterman, M. Aladjem, and I. Dinstein, "On the initialization of Sammon's nonlinear mapping," *Pattern analysis and applications*, vol. 3, pp. 61–68, 2000.
[49] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proc. 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, University of California Press, vol. 1, pp. 281–297, 1967.
[50] A. K. Jain, and R. C. Dubes, *Algorithms for clustering data*. Englewood Cliffs, New Jersey: Prentice Hall, 1988.
[51] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Comput. Surv.*, vol. 31, pp. 264–323, 1999.
[52] T. M. Cover, and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Information Theory*, vol. IT-13, no. 1, pp. 21–27, 1967.
[53] C. M. van der Walt, and E. Barnard, "Data characteristics that determine classifier performance", in *Proc. Sixteenth Annual Symposium of the Pattern Recognition,* Association of South Africa, pp.160–165, 2006.
[54] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification.* Wiley Interscience, 2nd ed., 2000, ch. 4.
[55] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math*., vol. 20, pp. 53–65, 1987.
[56] L. Kaufman, and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley Interscience, 1990.
[57] V. E. McGee, and W. T. Carleton, "Piecewise regression," *J. Am. Stat. Assoc.*, vol. 65, pp. 1109–1124, 1970.
[58] M. N. Karim, D. Hodge, and L. Simon, "Data-based modeling and analysis of bioprocesses. Some real experiences," *Biotechnol. Prog.*, vol. 19, pp. 1591–1605, 2003.
[59] W. S. Cleveland, E. H. Grosse, and W. M. Shyu, *Local regression models*. London: Chapman and Hall, J. M. Chambers, and T. J. Hastie, Eds., 1992, pp. 309–376.
[60] Y. Chen, G. Dong, J. Han, B. W. Wah, and J. Wang, "Multi-dimensional regression analysis of time-series data streams," *Proc. 28th Int. Conf. Very Large Data Bases*, Hongkong, China, pp. 323–334, 2002.

[61] Akhbardeh, A., Nikhil, Koskinen, P.E., Yli-Harja, O., Towards the Experimental Evaluation of Novel Supervised Fuzzy Adaptive Resonance Theory for Pattern Classification, *Pattern Recognition Letters* (2007), doi: 10.1016/j.patrec.2007.10.017
[62] Ramkrishna D, Amundson NR (2004) Mathematics in chemical engineering: a 50 year introspection. *AIChE J* 50:7-23
[63] G. Endo, T. Noike and J. Matsumoto, "Characteristics of cellulose and glucose decomposition in acidogenic phase of anaerobic digestion," *Proc. Soc. Civ. Engrs.*, vol. 325, pp. 61–68, 1982. (In Japanese).
[64] H. Q. Yu, Z. H. Hu, T. Q. Hong and G. W. Gu, "Performance of an anaerobic filter treating soybean processing wastewater with and without effluent recycle," *Process Biochem.*, vol. 38, pp. 507–513, 2002.
[65] N. Kataoka, A. Miya, and K. Kiriyama, "Studies on hydrogen production by continuous culture system of hydrogen-producing anaerobic bacteria," *Water Sci. Technol*., vol. 36, no. 6-7, pp. 41–47, 1997.
[66] C. C. Chen, and C.-Y. Lin, "Using sucrose as a substrate in an anaerobic hydrogen producing reactor," *Adv. Environ. Res.*, vol. 7, pp. 695–699, 2003.
[67] C.-Y. Lin, and C. H. Lay, "Carbon/nitrogen-ratio effect on fermentative hydrogen production by mixed microflora," *Int. J. Hydrogen Energy*, vol. 29, no. 1, pp. 41–45, 2004.
[68] C.-Y. Lin, and C. H. Lay, "Effects of carbonate and phosphate concentrations on hydrogen production using anaerobic sewage microflora," *Int. J. Hydrogen Energy*, vol. 29, no. 3, pp. 275–81, 2004.
[69] M. Dubois, K. A. Giles, J. K. Hamilton, P. A. Rebers, and F. Smith, "Colorimetric method for determination of sugars and related substances," *Anal. Chem.*, vol. 28, pp. 350–356, 1956.
[70] APHA. 1995. *Standard methods*. 19th Edition. American Public Health Association, Washington, DC.
[71] Koskinen PEP, Kaksonen AH and Puhakka JA (2007) The relationship between instability of H2 production and compositions of bacterial communities within a dark fermentation fluidized-bed bioreactor. *Biotechnol Bioeng* 97(4):742-758
[72] Hawkes, F.R.; Hussy, I.; Kyazze, G.; Dinsdale, R.; Hawkes, D. L. Continuous dark fermentative hydrogen production by mesophilic microflora: Principles and progress. *International Journal of Hydrogen Energy* 2007, 32, 172 – 184.