Journal of Biotechnology 147 (2010) 186–197

Contents lists available at ScienceDirect

Journal of Biotechnology

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

Mining manufacturing data for discovery of high productivity process characteristics

Salim Charaniya[a,1], Huong Le[a], Huzefa Rangwala[b], Keri Mills[c], Kevin Johnson[c], George Karypis[d], Wei-Shou Hu[a,*]

[a] Department of Chemical Engineering and Materials Science, University of Minnesota, 421 Washington Avenue SE, Minneapolis, MN 55455-0132, United States
[b] Department of Computer Science, George Mason University, 4400 University Drive, Fairfax, VA 22030, United States
[c] Genentech, Inc., 100 New Horizons Way, Vacaville, CA 95688, United States
[d] Department of Computer Science and Engineering, University of Minnesota, 200 Union Street SE, Minneapolis, MN 55455, United States

ARTICLE INFO

Article history:
Received 4 February 2010
Received in revised form 1 April 2010
Accepted 13 April 2010

Keywords:
Data mining
Multivariate data analysis
Support vector machine
Bioprocess
Cell culture
Manufacturing

ABSTRACT

Modern manufacturing facilities for bioproducts are highly automated with advanced process monitoring and data archiving systems. The time dynamics of hundreds of process parameters and outcome variables over a large number of production runs are archived in the data warehouse. This vast amount of data is a vital resource to comprehend the complex characteristics of bioprocesses and enhance production robustness. Cell culture process data from 108 'trains' comprising production as well as inoculum bioreactors from Genentech's manufacturing facility were investigated. Each run constitutes over one-hundred on-line and off-line temporal parameters. A kernel-based approach combined with a maximum margin-based support vector regression algorithm was used to integrate all the process parameters and develop predictive models for a key cell culture performance parameter. The model was also used to identify and rank process parameters according to their relevance in predicting process outcome. Evaluation of cell culture stage-specific models indicates that production performance can be reliably predicted days prior to harvest. Strong associations between several temporal parameters at various manufacturing stages and final process outcome were uncovered. This model-based data mining represents an important step forward in establishing a process data-driven knowledge discovery in bioprocesses. Implementation of this methodology on the manufacturing floor can facilitate a real-time decision making process and thereby improve the robustness of large scale bioprocesses.

1. Introduction

In the past decade, we have witnessed an explosion of data in almost every aspect of life. Increasingly such data are well organized and annotated in data warehouses affording the opportunities for mining for knowledge discovery. In many industrial sectors, including finance, retails and services, the data-driven approach has been widely used for discerning the trend of customer or market behavior (Usama and Ramasamy, 1996). Mining data warehouse has also attracted much attention in biotechnological sector partly because of the rapid expansion of genomics and other -omics based data. Recent increase in biologics manufacturing capacity also present an area of data mining that is yet to be explored.

Over two dozen new biologics have been licensed for therapeutic applications in the past few years (Aggarwal, 2009). This was accompanied by an expansion of manufacturing facilities around the world. These modern manufacturing facilities are highly automated in their operation and data acquisition. Hundreds of process parameters are routinely acquired and archived electronically, not only at the production scale, but also throughout cell expansion in the inoculum 'train'. Invariably fluctuations in process productivity and product quality occur in those production facilities. Such fluctuations or variations may exist in the same plant over time, or in plants at different locations for the same product. Understanding the root of such variations and enhancing process robustness will have major economic implications for the product. Mining bioprocess data to identify parameters which may be related to process fluctuations holds much potential for enhancing the productivity and process consistency.

Several studies in the past have employed a variety of techniques to explore bioprocess data. Principal component analysis (PCA), partial least-squares (PLS) and unsupervised clustering have been proposed to analyze and monitor bioprocesses (Bachinger

* Corresponding author. Tel.: +1 61 2 626 7630; fax: +1 61 2 626 7246.
E-mail address: acre@cems.umn.edu (W.-S. Hu).
[1] Present address: Genentech, Inc., 1 Antibody Way, Oceanside, CA 92056, United States.

0168-1656/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jbiotec.2010.04.005

et al., 2000a; Kamimura et al., 2000; Kirdar et al., 2007). A decision tree-based classification approach was proposed to identify the process trends that best differentiate runs with high and low productivity (Bakshi and Stephanopoulos, 1994; Stephanopoulos et al., 1997). Artificial neural network (ANN) is also a popularly used tool to model the non-linear interactions in temporal process data (Bachinger et al., 2000b; Coleman and Block, 2006; Coleman et al., 2003; Glassey et al., 1994a,b; Huang et al., 2002; Vlassides et al., 2001).

Despite previous attempts, mining vast volumes of production-scale process data and on-line implementation of such schemes remain arduous. Bioprocess datasets are unique in their heterogeneity; the frequency of measurement varies widely among on-line and off-line parameters. Also, the types of data include both discrete (e.g. valve settings as ON/OFF) and continuous values (for a recent review, see (Charaniya et al., 2008)). In addition to temporal measurements of viability, viable cell densities, consumption and production rates of key nutrients and metabolites, a plethora of process-associated parameters are commonly recorded. These include not only temporal records of physical parameters, such as temperature, dissolved oxygen concentration, aeration rate and reactor mass, but also those involved in process control loops for physical parameters, such as a battery of valves for gas flow, base addition and nutrient feeding. Additionally, the records of raw materials with respect to source, lots, quantity as well as the timestamps for preparation and material addition are maintained. The complexities associated with the enormity and the unique characteristics of bioprocess data present substantial challenges as well as opportunities for process data mining.

In a recent report, we outlined a systematic procedure for analyzing bioprocess data (Charaniya et al., 2008). Bioprocess datasets often require a preprocessing step involving transformation, normalization, and computation of missing values. A dimensionality reduction step is often used subsequently to identify a subset of process features that are more informative for data mining. The data mining step involves application of descriptive (e.g., frequent pattern discovery, clustering) and predictive (e.g. classification, regression) pattern recognition methods to discover significant trends in process data. These models allow one to comprehend how different process parameters and their interactions affect process outcome. The identified trends or models can be interpreted by process experts to gain further insights for process improvement.

Support vector machines (SVM) are a class of predictive machine learning algorithms motivated by the Vapnik–Chervonenkis theory that laid the foundations of the principle of structural risk minimization (SRM) (Vapnik, 1998a,b). Originally formulated for binary classification, SVM models identify a linear *decision boundary* that separates objects (e.g. process runs) from the two distinct classes with maximum distance (called *margin*). Here, each object is characterized by a set of features, (e.g. process features extracted from the process data of a run). Also, non-linear SVM models can be constructed by using kernel transformation functions (Muller et al., 2001). Due to their good generalizability (i.e., predicting the class/outcome of objects not used for model construction) and scalability with respect to the dimensionality of the feature space, SVMs have gained immense popularity as a pattern classification tool in several data-intensive fields such as computational biology (Ben-Hur et al., 2008).

In this study, we analyzed a vast volume of on-line as well as off-line cell culture process data acquired during production runs at the Genentech's recombinant protein manufacturing facility at Vacaville, CA. Process data from 108 runs were scrutinized to investigate the variation in process outcome and to identify the distinguishing characteristics of high productivity processes. After data preprocessing, we employed a kernel-based SVM learning technique to establish an adaptable data mining framework. This framework integrates off-line and on-line temporal process data to construct support vector regression models that can predict process outcome and identify critical parameters that shed insights on process productivity.

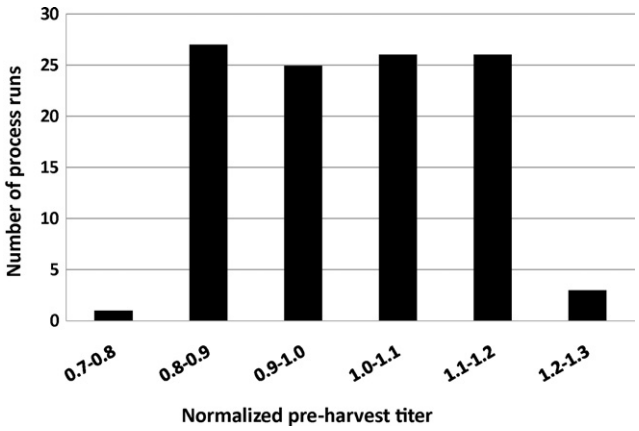

Fig. 1. Histogram of the normalized pre-harvest titer of 108 production runs.

2. Methods

2.1. Data preprocessing

The culture bioreactors from which the data were analyzed in this study were located at Genentech's Vacaville facility comprising scales ranging from 20 L to 12,000 L. The recombinant mammalian cells were expanded from the cell bank to the 20 L scale and then step-wise scaled up at 80 L, 400 L, and 2000 L. A recombinant Chinese Hamster ovary (CHO) cells producing immunoglobulin G were cultivated for approximately 75 h in the bioreactors at each scale before they were inoculated in the production scale (12,000 L) bioreactors where they were cultured for approximately 11 days. In this study, process data obtained from 108 runs was analyzed. Each run comprises cell culture process data from 80 L, 400 L, 2000 L, and 12,000 L scale bioreactors.

For every run, the antibody concentration (called titer) was measured at the end of the cell culture process and normalized to an average titer of 1.0. The normalized titer of the 108 runs distributes over a range (Fig. 1).

2.1.1. On-line parameters

The manufacturing facility is equipped with automated control and data logging systems whereby acquired process data are recorded and archived on-line electronically. Over 130 parameters were acquired on-line at each of the three inoculum scales and the production scale bioreactors. The on-line parameters include control parameters and control action parameters. The former category includes parameters such as dissolved oxygen (DO), pH, and vessel temperature that are controlled at specific levels (e.g., vessel temperature at 37 °C), whereas the latter category includes parameters such as controller responses, the sparge rates of air and oxygen to control DO, and the rates of base addition and carbon dioxide sparge to control pH. Other important parameters such as vessel volume and overlay gas flow rates are also acquired on-line. The volumetric oxygen uptake rate (OUR) is estimated approximately every 4 h, whereas all other on-line parameters are acquired almost continuously (once every few seconds) over the entire duration of the run that lasts several days. In addition to these parameters whose values are continuous, there are 'discrete' parameters such as the state of different valves, which is often binary (OFF/ON state). These valves control different ports for addition of inoculum, media, base,

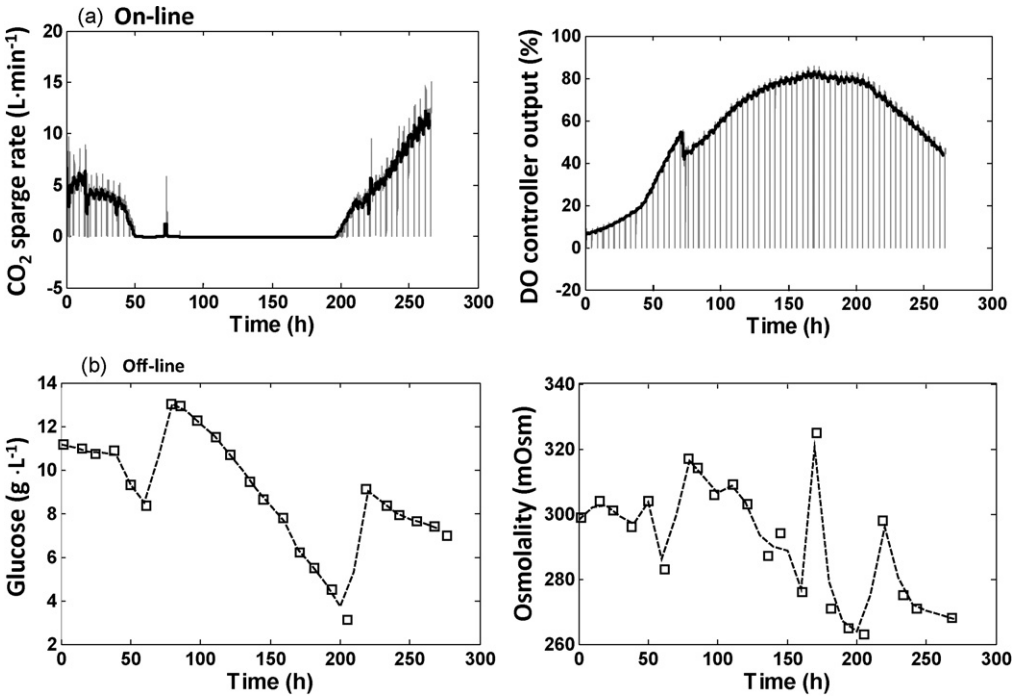

Fig. 2. Preprocessing of cell culture process data. (a) On-line parameters were preprocessed using a moving window average method (see Section 2). The panels display the temporal profiles of carbon dioxide sparge rate (left panel) and DO controller output (right panel) from a 12,000 L fed-batch culture. ( ) Measured, ( ) preprocessed. (b) Off-line parameters were preprocessed using a linear interpolation scheme. The panels show the temporal profiles of glucose concentration (left panel) and medium osmolality (right panel) from a 12,000 L fed-batch culture. (□) Measured, (—) interpolated.

antifoam, and gas sparging among others. At least 40 parameters related to states of different valves are recorded at each of the four bioreactor scales.

On-line data were preprocessed using a moving window average method. A time window of 100 min was selected. At every time point, a parameter value was approximated as the average of all the measurements for the parameter within the time window. For instance, the processed value at time t is the average of measurements at time t, $t+1$, $t+2$, …, $t+99$ min. The raw and the preprocessed temporal profiles of CO_2 sparge rate and DO controller output at the 12,000 L scale of one run are shown in Fig. 2a as examples. The preprocessed profile delineates the temporal patterns of these parameters without the disturbances at the local timescales.

2.1.2. Off-line parameters

A number of parameters related to nutrient consumption and metabolite production are measured off-line by periodic withdrawal of samples from the bioreactors (Table 1). The parameters include physical and state parameters, chemical parameters, and physiological parameters. A total of 12 parameters at the production scale and 11 parameters at each of the three inoculum scales were measured periodically. Due to the differences in sampling frequencies of the off-line parameters, all off-line measurements were preprocessed using a linear interpolation method. Fig. 2b shows the temporal profiles of glucose concentration and medium osmolality in the production bioreactor for one run. The aggregation of off-line and on-line data from the inoculum train and the production bioreactor exceeds more than one million data points for a single production run.

2.1.3. Other parameters

Three parameters related to cell characteristics were also selected for comparative analysis. These are cell bank, cell ampoule, and cell age. Different working cell banks (WCBs) were used during the course of the 108 runs. Each WCB comprises multiple cell ampoules, each of which is thawed and undergoes multiple passages to initiate several process runs. Thus, each process run is associated with a cell ampoule from a WCB. An additional parameter is the cell age which is related to the time period during which the cells were sustained in the 20 L seed bioreactor. In addition to these parameters, different raw material lots of hydrolysate used in the cell culture medium at the production scale were also examined.

Table 1
Summary of process parameters at different bioreactor scales.

Off-line parameters	On-line parameters
Physical and state parameters	*Controlled parameters*
Dissolved carbon dioxide	Dissolved oxygen (primary probe)
Dissolved oxygen	Dissolved oxygen (secondary probe)
pH (off-line)	Vessel temperature
Chemical parameters	pH (on-line)
Lactic acid concentration	Jacket temperature
Glucose concentration	*Control action parameters*
Sodium ion concentration	Dissolved oxygen (DO) controller output
Ammonium ion concentration	Air sparge rate
Osmolality	Air sparge set point
Physiological parameters	Total air sparged
Viable cell density	Oxygen sparge rate
Viability	Total oxygen sparged
Packed cell volume	pH controller output
Integral of packed cell volume[a]	Total base added
Other parameters	CO_2 sparge rate
Cell bank	Total CO_2 sparged
Cell ampoule	Total gas sparged
Cell age	*Others*
Hydrolysate material lot[a]	Oxygen uptake rate
	Reactor weight
	Overlay flowrate
	Exhaust valve pressure
	Backpressure

[a] Only at 12,000 L scale.

2.2. Estimation of similarity between runs

A crucial aspect of our approach is to compare the 'likeness' of any two runs based on the process parameter profiles. In a recent report, we proposed a two-step method to compute the similarity between two runs (say run 1 and run 2) (Charaniya et al., 2008). In the first step, individual parameters (e.g. osmolality profile, hydrolysate material lots, etc) from run 1 are compared with the corresponding parameters in run 2, and a similarity score is computed for each parameter. In the second step, all individual parameter-wise similarity scores are integrated to estimate the overall similarity between the two runs.

2.2.1. Similarity between runs for individual parameters

The temporal profile of a process parameter (p) was compared between any two runs (denoted by i and j) using the Euclidean distance metric (d_{ij}^p).

$$d_{ij}^p = \left\| p_i - p_j \right\| = \sqrt{\sum_{k=1}^{l} (p_{ik} - p_{jk})^2}, \quad (1)$$

where p_{ik} corresponds to the measured value of the parameter at time point k in run i.

For n different runs, the parameter profiles ($p_1, p_2, \ldots, p_n$) were compared in a pairwise manner. The resulting Euclidean distances were scaled between 0 and 5, where 0 corresponds to the highest similarity, and 5 corresponds to the lowest similarity between two profiles. The Euclidean distance metric (d_{ij}^p) was translated into a similarity metric (s_{ij}^p) using an exponential transformation; i.e.

$$s_{ij}^p = \exp(-d_{ij}^p). \quad (2)$$

The similarity metric ranges between 0.01 ($\exp(-5)$) for dissimilar profiles and 1.00 ($\exp(0)$) for identical profiles. All the pairwise estimates of the similarity of a parameter profile across different runs comprise a similarity matrix for that parameter. The similarity matrix is symmetric, and positive semidefinite (i.e., all the eigenvalues of the matrix are non-negative), thus satisfying the Mercer's theorem. The similarity matrix is, therefore, a valid Mercer kernel.

Cell ampoule and cell bank used for different runs were compared using a binary metric, i.e., the similarity score is 1 if the runs used the same cell source and 0 otherwise. Cell age between two runs was compared by estimating the absolute value of the difference in age (in days) between the two runs. This difference, after being scaled between 0 and 5, was translated into a similarity metric using an exponential transformation (Eq. (2)). Lastly, 12 different hydrolysate material lots were used in the culture medium at the production scale for the 108 runs. For each run, a 12-dimensional vector was created where each dimension represents the fractional amount of a particular lot used in that run. The runs were thereafter compared in a pairwise manner using the Pearson's correlation coefficient and the similarity values were scaled between 0 and 1.

2.2.2. Overall similarity between different runs

The likeness between two runs (i and j) was computed by a weighted linear combination of the similarity between individual parameters. For example, for three parameters (p, q, r) measured in runs i and j, the overall similarity is estimated as:

$$s_{ij} = w_p s_{ij}^p + w_q s_{ij}^q + w_r s_{ij}^r, \quad (3)$$

where w_p, w_q, w_r are the weighting factors for parameters p, q, and r respectively.

2.3. Estimation of parameter weight

A weight was assigned to every parameter by comparing the similarity of that parameter profile between any two runs with the difference in the outcome of these two runs. Final titer was used as a measure of process outcome. For every parameter, all possible combinations of two runs were compared. The difference in their final titers was correlated to the similarity between the temporal profiles of the parameter using the Spearman's rank correlation coefficient (*rho*). The weights for individual parameters were obtained by scaling *rho* such that the sum of all the weights is equal to one.

2.4. Estimation of parameter interdependency

All the pairwise similarity scores from the similarity matrix of a parameter were re-arranged as a one-dimensional vector of similarity scores. Thus, each parameter is represented by a similarity vector that comprises the similarity scores between its profiles in all possible pairwise combinations of runs. These similarity score vectors for all parameters were pairwise compared using the absolute value of the Pearson's correlation coefficient. All the parameters were clustered using a hierarchical clustering scheme in which the inter-cluster similarity was calculated as the weighted average of the Pearson's correlation between all pairs of parameters from the two clusters (weighted average linkage, or WPGMA). The resulting dendrograms were pruned at a Pearson's correlation threshold of 0.6 to obtain several clusters of parameters. For each multi-parameter cluster, the parameter with the highest similarity to the cluster centroid was selected to represent that cluster.

2.5. Supervised machine learning

2.5.1. Support vector regression (SVR)

Support vector regression implements the support vector algorithm for estimating a regression function for the outcome variable (denoted as y). A set of n training runs can be denoted as $(x_i, y_i)\ \forall\ i = 1, 2, \ldots, n$, where $x_i \in \mathbb{R}^d$ is the space of all input parameters, and y_i is the outcome of the ith run. A support vector regression (SVR) model seeks to identify a regression function $f(x) = \mathrm{w} \cdot \mathrm{x} + b$ that minimizes the difference (i.e., the error) between the true process outcome (y_i) and the model-predicted outcome ($f(x_i)$). A ν-SVR algorithm was employed to estimate the regression function (Scholkopf et al., 2000). For each run (x_i, y_i), an error $|y_i - f(x_i)|$ of up to ε is considered acceptable. Differences exceeding ε are penalized by a slack variable (ξ_i or ξ_i') and an *a priori* chosen cost function (C). The parameters (w,b) of the regression function are obtained by solving a constrained optimization problem:

$$\min_{w,b} \left\{ \frac{1}{2}\|w\|^2 + C\left(\nu\varepsilon + \frac{1}{n}\sum_{i=1}^{n} (\xi_i + \xi_i') \right) \right\}, \quad (5)$$

subject to the following inequality constraints ($\forall i = 1, 2, \ldots, n$)

$$\begin{aligned} &(wx_i + b) - y_i \le \varepsilon + \xi_i. \\ &y_i - (wx_i + b) \le \varepsilon + \xi_i' \\ &\xi_i, \xi_i' \ge 0\ ;\ \varepsilon \ge 0 \end{aligned} \quad (6)$$

The ν-SVR algorithm seeks to minimize the error ε. The parameter ν is a non-negative constant that determines the balance between the complexity of the model and the extent of the error ε. LIBSVM (Chang and Lin, 2001), an implementation of ν-SVR in C, was used for training and validation of the SVR models. The default value of $\nu = 0.5$ was used.

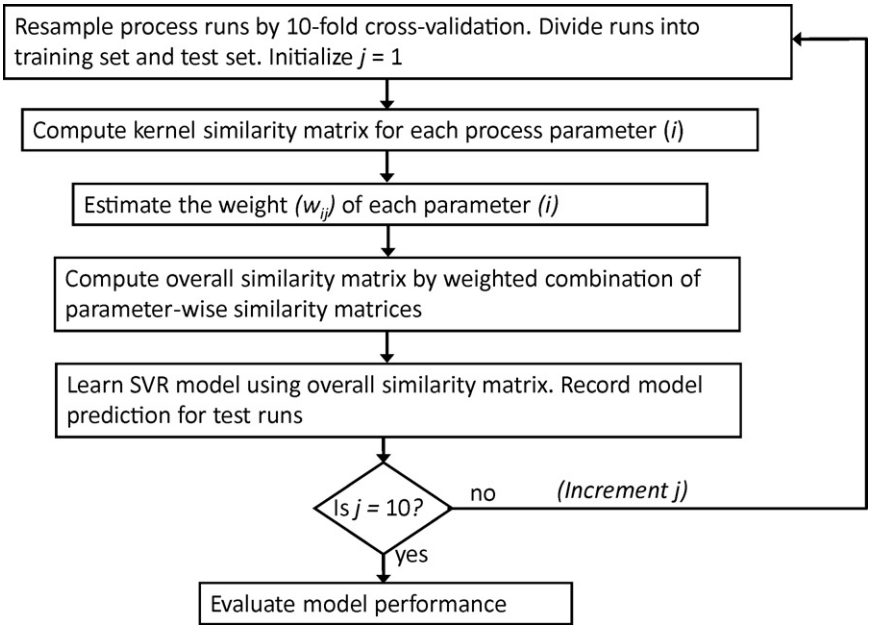

Fig. 3. Flow diagram of the proposed methodology for knowledge discovery in manufacturing cell culture process data.

2.5.2. Model training and evaluation

v-SVR models were trained using a dataset of 108 runs (n = 108). Supervised learning can, however, result in overfitting of the model. An overfitted model can successfully predict the final titer of the training runs, but the generalization error of the model is high (i.e.,its ability to predict the final titer of test runs, which were not used for model training, is poor). A 10-fold cross-validation approach was used to assess the generalizability of the model (Fig. 3). The training dataset was divided into ten groups. In each iteration, nine of the ten groups comprise the training set, and parameter weights were evaluated based on the training set only. The parameter weights and profile similarities were used to construct an SVR model based on the training runs. The model was then used to predict the final titer of the runs in the 10th test group.

The above procedure was reiterated ten times. In each iteration, a different group was used for testing. Parameter weights were estimated and an SVR model was constructed from the training runs. Model performance was assessed by comparing the model-predicted titers and the actual titers of the test runs using the Pearson's correlation coefficient and the root mean square error (RMSE). RMSE is a measure of the average error in predicting the final titer of a run:

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n}(y_i - f(\mathbf{x}_i))^2}, \quad (7)$$

where y_i and $f(x_i)$ are the actual and the model-predicted titers of run i, respectively.

3. Results

3.1. Selection and preprocessing of bioprocess data

The parameters acquired on-line have different levels of relevance to the process outcome. They also vary in their richness in information pertinent to process characteristics. A preliminary survey was performed using process data from 30 runs to identify a subset of on-line parameters for further analysis. This subset of parameters consists of 21 on-line parameters at the 12,000 L scale, and 20 parameters at each of the three inoculum scales (80 L, 400 L and 2000 L). In order to reduce noise at local timescales and also dampen the effect of the local discontinuities observed due to periodic interventions, all the on-line parameters were preprocessed using a moving window average method. Off-line parameters were also preprocessed by linear interpolation (see Section 2).

The resulting preprocessed data comprises a total of 126 temporal parameters: 33 (21 on-line, 12 off-line) at the 12,000 L scale, and 31 (20 on-line, 11 off-line) parameters at each of the three inoculum scales. The preprocessed parameter profiles were compared to estimate the similarities and differences in their temporal patterns across different runs. In addition to these temporal parameters, three parameters related to cell source (cell bank, cell ampoule and cell age) and the hydrolysate material lots used at the production scale bioreactors were also compared across different runs.

3.2. Kernel transformation and comparison of process runs

3.2.1. Parameter profile-based comparison of process runs

A Euclidean distance metric was used for comparison of time-dependent parameters. The Euclidean distance was converted to a similarity score by an exponential kernel transformation (see Section 2). This similarity comparison was repeated for all pairwise combinations of runs. For example, the osmolality profile from the 12,000 L scale of every process run was compared with the osmolality profiles at the 12,000 L scale of each of the other 107 runs in a pair-wise manner. The results comprise a 108 × 108 kernel similarity matrix for osmolality. Similarly, kernel similarity matrices were generated for all the parameters at the four different bioreactor scales.

In addition, a temporal alignment was performed for each off-line parameter across all pairwise combinations of process runs. Parameter profiles for a pair of runs were allowed to shift within a window of ±10 h such that the Euclidean distance between them is minimized. Kernel similarity matrices were re-calculated and compared to the corresponding matrices obtained without any time shift. An average Pearson's correlation of 0.88 was observed between the similarity matrices obtained before and after time alignment indicating that there is no significant temporal shift between the runs. Therefore, kernel similarity matrices obtained without any time alignment were used for subsequent analysis.

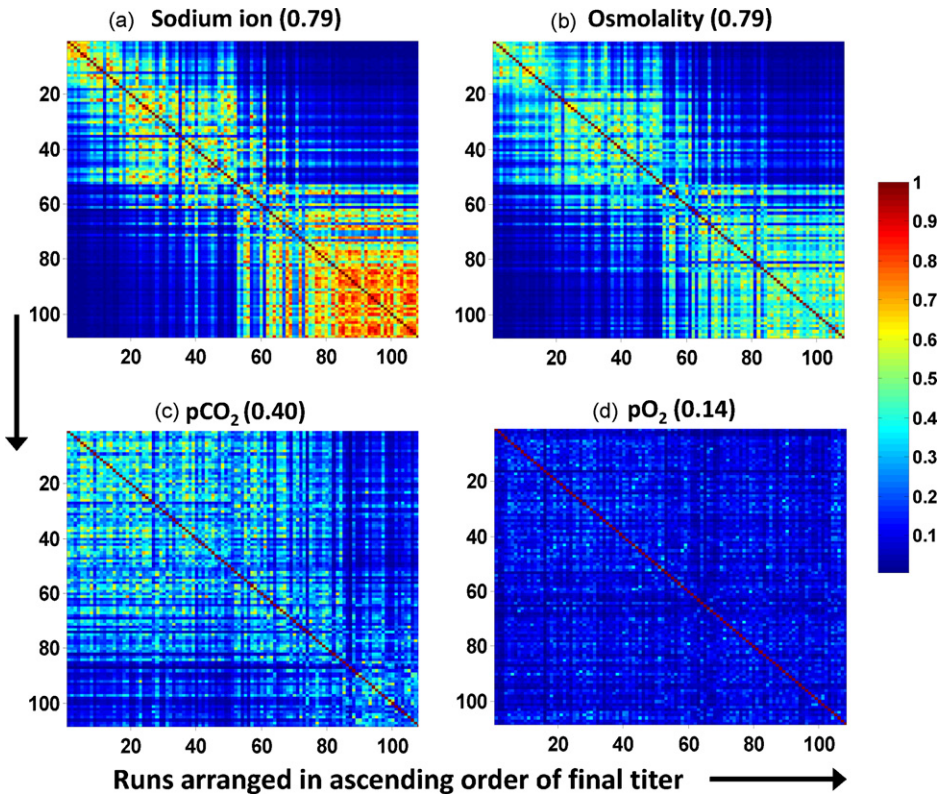

Fig. 4. A differential weighting scheme for process parameters. The kernel similarity matrices for four process parameters acquired at the 12,000 L scale are shown. The parameters are (a) sodium ion concentration, (b) medium osmolality, (c) dissolved carbon dioxide (p_{CO_2}), (d) Dissolved oxygen (p_{O_2}). Each element (i,j) of a parameter kernel matrix represents the similarity (s_{ij}) between the temporal profiles of that parameter between two runs (run i and run j). The similarity score ranges from 0.01 (blue) for dissimilar profiles to 1 (red) for identical profiles. Note that the matrix is symmetric, i.e., $s_{ij} = s_{ji}$, and all diagonal values are 1, i.e., $s_{ii} = 1$. The runs are arranged in increasing order of the normalized final titer—run 1 has the minimum titer and run 108 has the maximum titer. A correlation between decreasing parameter profile similarity and increasing titer difference between runs is observable as a red-to-blue gradient. The number in parenthesis is the absolute value of the Spearman's rank correlation coefficient for each of the four parameters. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

The integration of similarity matrices of different parameters involves aggregation of those parameter-wise similarity scores into an overall estimate of the likeness of two process runs. Here, we adopted a scheme of weighted combination of these parameter-wise similarity scores such that critical process parameters have greater contribution to the overall similarity between two runs.

3.2.2. Productivity-based approach for parameter weighting

The scheme for assigning the weights of different parameters to determine the similarity between a pair of runs can significantly affect the analysis. A simple, yet proficient approach was used to assign higher weights to the process parameters that correlate with the outcome variable (i.e., product titer). This can be illustrated with an example of four process runs (runs 1, 2, 3, and 4) with the normalized final product titers of 0.8, 0.9, 1.0, and 1.1, respectively. The temporal profile of a parameter in run 1 is pair-wise compared with the parameter profiles in the other three runs. Consider a case where all the three similarity scores are 0.9 (on a 0–1 scale). In this case, the parameter has a comparable profile across all the four runs. This indicates that the parameter does not provide much information for deciphering the differences in the outcome of the four runs. In contrast, consider a case where the three similarity scores (between runs 1–2, 1–3, and 1–4) are 0.9, 0.6, and 0.4, respectively. Here, the *increasing* titer-difference between runs 1–2, 1–3, and 1–4 correlates with the *decreasing* similarity between the parameter profiles. The trend of the similarity scores for a given parameter among different pairs of runs can be quantified using the Spearman's rank correlation (*rho*), a non-linear measure for assessing the correlation between two variables. In the above example, the *rho* between titer-differences (0.1, 0.2, and 0.3) and similarity scores (0.9, 0.6, and 0.4) for the parameter is −1.0, indicating that the parameter profile can discriminate between runs with different titers.

For each parameter, the trend between the similarity score (between every pair of runs) and the titer-difference (between the two runs) was assessed by the *rho* metric across all pairs of runs. Fig. 4 shows the similarity matrices for four process parameters acquired at the 12,000 L scale. The 108 runs are arranged in an increasing order of the final titer, such that run 1 has the lowest titer and run 108 has the highest titer. Every element of a parameter's similarity matrix represents the similarity score of the temporal profiles of that parameter between two runs. As one moves away from the diagonal, the titer-difference between the runs increases. For parameters such as sodium ion concentration and medium osmolality, elements at farther distances from the diagonal also exhibit lower similarity scores. Such a trend is not seen for dissolved oxygen (p_{O_2}) (Fig. 4). Thus, the *rho* metric, which quantifies this trend, is a measure of the degree of importance of every parameter. We assigned weighting factors proportional to *rho* such that parameters with high negative *rho* have greater weights.

3.2.3. Integration of all process parameters

The weighted sum of similarity scores of all parameters is taken as the overall similarity between any two runs. The overall similarity was estimated for every pair of runs to obtain a 'fused' kernel matrix. Using the comparative information in this kernel matrix, we constructed models to predict the final outcome of different runs.

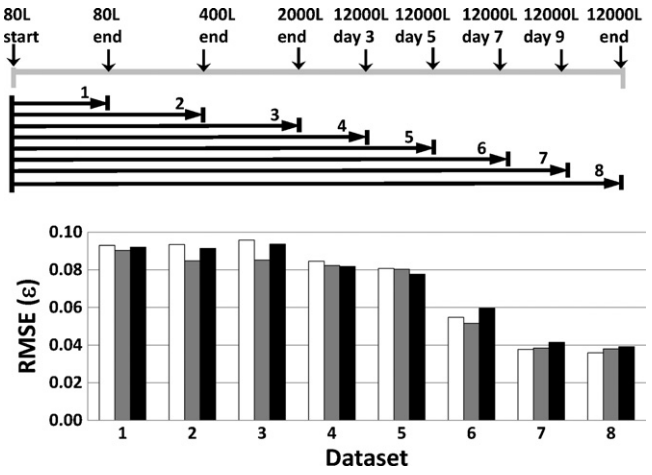

Fig. 5. Evaluation of SVR model performance. Top panel: The process data from the inoculum bioreactor scales (80 L, 400 L and 2000 L) and the production bioreactor scale (12,000 L) was divided into eight datasets in order to analyze process data in a stage-wise cumulative manner. The timescale of each dataset is shown in the process timeline. For each dataset, a 10-fold cross-validation procedure was used to assess the performance of the SVR models. Bottom panel: Root mean square error (ε) of the SVR models constructed using (□). All parameters: A weighted combination of all process parameters was used for SVR model construction; (■) Top 10 parameters: During each iteration of cross-validation, the top 10 parameters were selected from the training dataset based on the parameter weighting factors (*rho*). SVR model was constructed from the top 10 parameters and the performance was evaluated on the test dataset; (■) Representative parameters: Hierarchical clustering was performed to compare parameters (based on their similarity matrices) to identify correlated parameters (Fig. 7 and Section 2). The parameter closest to the cluster centroid was identified to represent each cluster of parameters. SVR models were constructed by a weighted combination of the representative parameters.

3.3. Predictive data mining using support vector regression

3.3.1. Process datasets

SVR was used to construct models for predicting process outcome, the final product titer. Several SVR models were formulated to investigate the progression of the production trains by gradually incrementing the dimensionality of the process dataset (Fig. 5a, top panel). Thus, the first dataset comprises 20 on-line and 11 off-line process parameters from the 80 L inoculum bioreactors only. Additionally, the three cell source-related parameters were included in this dataset. A fused kernel matrix was obtained by computing the overall similarity for every pair of runs based on the first dataset alone. Using this kernel matrix, an SVR model was constructed to predict the final titer of the run. Note that the final titer was determined upon the completion of the 12,000 L run or approximately 17 days from the end of the 80 L run. Thus, SVR models based on the first dataset employs data measured very early in the inoculum train to predict the final process outcome.

The second dataset combines process data from the 80 L and 400 L inoculum bioreactors and the third dataset cumulates data from all the three inoculum scales. Similarly, the subsequent four datasets (datasets 4, 5, 6, and 7) incorporate process data up to day 3, day 5, day 7, and day 9 of the 12,000 L bioreactors, respectively. Lastly, the eighth dataset integrates all the process data from the four scales. This cumulative organization of process data allows comparison of the predictability of the SVR models at various stages of the production train. In addition, the stage-wise comparison is advantageous for identifying critical process parameters at each production stage.

3.3.2. SVR models for predicting process outcome

As described in Section 2, a 10-fold cross-validation scheme was used to evaluate and compare the SVR models constructed for all the eight datasets described above. A two-step grid search with 10-fold cross-validation was also performed to obtain the optimal cost functions. A coarse grid comprising values of 10^{-6}, 10^{-3}, 0.01, 0.05, 0.1, 0.5, 1, 10, 100, and 10^{6} was used in the first step, which identified the optimal region between 0.1 and 1. A finer search was subsequently performed in the region from 0.1 to 1 to obtain the optimal value of the cost function as 0.1, which was used in constructing all the SVR models.

Random predictors were also used to assess the significance of the SVR models. Based on 10000 simulations of randomized titer prediction, the Pearson's correlation (between actual and predicted titer) is expectedly zero and the root mean square error (RMSE) is 0.176. In contrast, the SVR models based solely on process data from the 80 L inoculum bioreactors (dataset 1) has significantly high predictability with the Pearson's correlation (r) and the RMSE (ε) of 0.619 and 0.093, respectively (Figs. 5 and 6). This indicates that, more than two weeks before process completion, the final titer can be predicted with noticeably better accuracy compared to a random predictor. Incorporation of process data from the 400 L inoculum bioreactors (dataset 2) results in a marginal increase in predictability of the SVR models with the evaluation metrics, r and ε, of 0.624 and 0.093, respectively (Fig. 5). Similarly, inclusion of data from the 2000 L inoculum bioreactors (dataset 3) did not result in any improvement in predictability ($r = 0.609$, $\varepsilon = 0.096$).

A marked improvement in model predictability is observed when process data from the first three days of the 12,000 L production scale bioreactors is included (dataset 4). The evaluation metrics, r and ε, improve to 0.734 and 0.085, respectively (Figs. 5 and 6). This trend of increasing predictability is conspicuous for datasets 5–8 where process data from additional days of the 12,000 L bioreactors is added sequentially. Thus, by the 7th day post-inoculation of the 12,000 L bioreactors (dataset 6), the final titer can be predicted with very high accuracy ($r = 0.909$, $\varepsilon = 0.055$) (Figs. 5 and 6). Lastly, the SVR model for dataset 7 has the best performance ($r = 0.952$, $\varepsilon = 0.038$), evincing that the final titer can be predicted very accurately by day 9 in the 12,000 L bioreactors.

The contribution of every process parameter in these SVR models is determined by a differential weighting scheme (described earlier). However, regardless of the degree of significance, all the process parameters contribute to model formulation, resulting in a high dimensionality. The detrimental effects of this 'curse of dimensionality' on data mining methods are well-known (Beyer et al., 1999). To alleviate this effect, the differential weighting scheme was used to reduce data dimensionality by pre-selecting a subset of top-weighting process parameters. SVR models were thereafter formulated using only this subset of parameters. For each of the eight process datasets, the top 10 process parameters were selected by the weighting scheme and SVR models were constructed by combining the kernel similarity matrices of the selected parameters only. Caution was exercised to avoid 'selection bias' by performing the top 10 parameter selection only on the training runs (without the inclusion of the test runs) (Ambroise and McLachlan, 2002). For datasets 1–3, SVR models based on the top 10 parameters exhibit a reduction in the RMSE (and an increase in the Pearson's correlation), indicating that outcome predictability is enhanced (Fig. 5). For example, in dataset 2, which includes process data from the 80 L and 400 L inoculum bioreactors, the RMSE is reduced by 9% from 0.093 for the SVR model with all parameters to 0.085 for the SVR model with the top 10 parameters. Also, the Pearson's correlation between the actual titer and the model-predicted titer increased from 0.624 to 0.693. However, for the subsequent datasets (4–8), differential selection of the top 10 parameters results in no significant change (datasets 6–8) or a decrease in predictability (datasets 4 and 5). Nonetheless, it is noteworthy that during the initial stages of the run in the inoculum train, a gain in predictability can be achieved by pruning the number of process parameters, thereby reducing the dimensionality of the dataset.

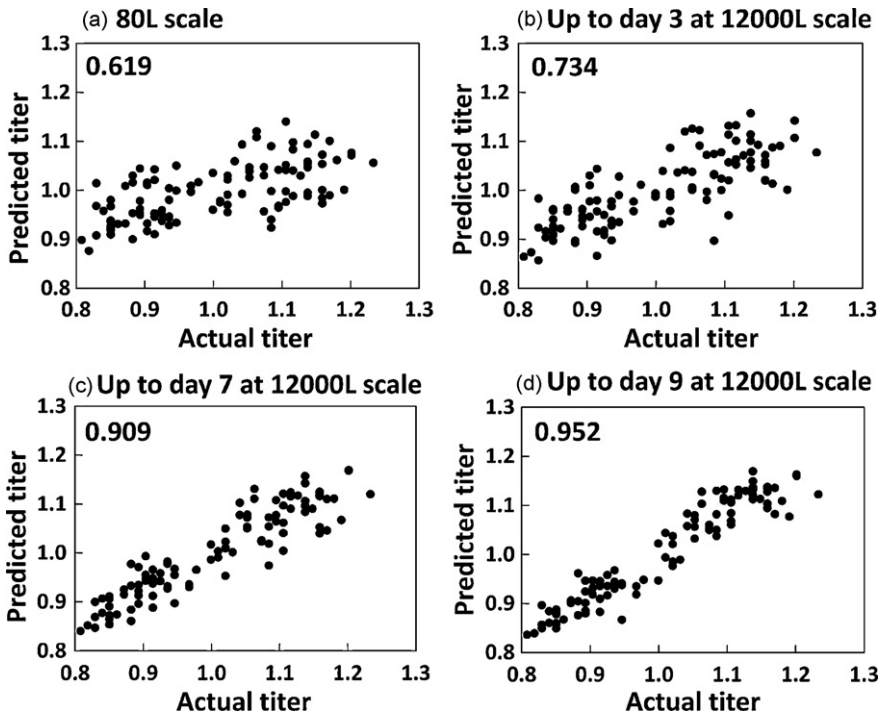

Fig. 6. SVR model performance at different bioreactor stages of the production train. (a) At the 80 L scale. (b) Up to day 3 at the 12,000 L scale. (c) Up to day 7 at the 12,000 L scale. (d) Up to day 9 at the 12,000 L scale. The Pearson's correlation coefficient (r) between the actual (normalized) titers and the model-predicted titers are shown in each panel.

3.4. Dimensionality reduction using similarity matrix-based parameter comparison

The consistent co-occurrence of several process parameters highly correlated with the final titer across different stages of the production phase (12,000 L) suggested a possible interdependency between them. Several process parameters are likely to be mutually related. For example, an increase in lactate production is often correlated to increased glucose consumption. Using individual parameter similarity matrices, the correlations between different process parameters were examined. A hierarchical clustering algorithm was used to group the correlated process parameters into a small number of clusters. Clustering was performed independently for each of the eight datasets (see Section 2). The clustering result for all process parameters in dataset 7 (up to day 9 at the production scale) is shown (Fig. 7). Osmolality, sodium ion concentration, total base added, lactic acid concentration and pH controller output are grouped into one cluster, indicating that the profiles of these parameters are correlated. Several parameters associated with different sparge rates, such as oxygen sparge rate, air sparge rate, and total gas sparged, also form a cluster. The key features provided by the correlated parameters within a cluster can be obtained by choosing a representative parameter from that cluster. For instance, total base added, which has the highest similarity to its cluster centroid, was chosen to represent all the parameters in that cluster.

Hierarchical clustering resulted in a smaller subset of relatively independent process parameters. SVR models were constructed using these parameter subsets and their predictability was compared with all-parameter models. For all datasets, SVR models constructed using parameter subsets exhibit performance comparable to the models with all parameters (Fig. 5). For example, for dataset 7, the 92-parameter subset model has the evaluation metrics, r and ε, as 0.946 and 0.041, respectively, compared to the 130-parmeter model ($r = 0.952$ and $\varepsilon = 0.038$). Thus, despite a 29% reduction in the number of parameters, the representative parameter subsets retain the essential process features resulting in SVR models with high predictability.

3.5. Stage-specific identification of critical process parameters

3.5.1. Weight-based assessment of process parameters

The Spearman's rank correlation (*rho*) used to assign the weight is indicative of the importance of each parameter in distinguishing between high and low titer runs. Recall that a parameter with a negative *rho* (and therefore a higher weight) correlates with deviations in process outcome. Among the 130 parameters, 97 (75%) have a *rho* less than 0.1. Further, among these 97 parameters, more than 69% have a *rho* less than 0.025, indicating negligible correlation with the final titer. Thirty-three parameters have a *rho* greater than 0.1, among which only ten have a strong correlation (*rho* > 0.5), highlighting that less than 8% of all process parameters are strongly correlated with the final titer.

The *rho* metric for process parameters at the 12,000 L scale for datasets 4, 5, 6, 7, and 8 were examined. Only those parameters with *rho* > 0.1 in at least one of the five datasets are shown (Fig. 8). A subset of parameters comprising reactor weight, on-line pH, oxygen sparge rate, DO controller output, and viability exhibit a strong increase in *rho* at day 9 (dataset 6) compared to the previous days (day 3, 5, and 7). Osmolality and reactor weight (i.e., load cell weight) profiles at the 12,000 L scale for all 108 runs are shown in Fig. 9a. It is evident that during the first 160 h, the profiles for these two parameters are similar between the high and low titer runs. A conspicuous difference between the two classes (high and low runs) emerges after 160 h. For the runs with low titer, there is a discernible increase in medium osmolality and an increase in reactor weight which corresponds to the increasing base and glucose addition for runs with low titer. Thus, the model is successful in discerning the temporal effects of the process parameters.

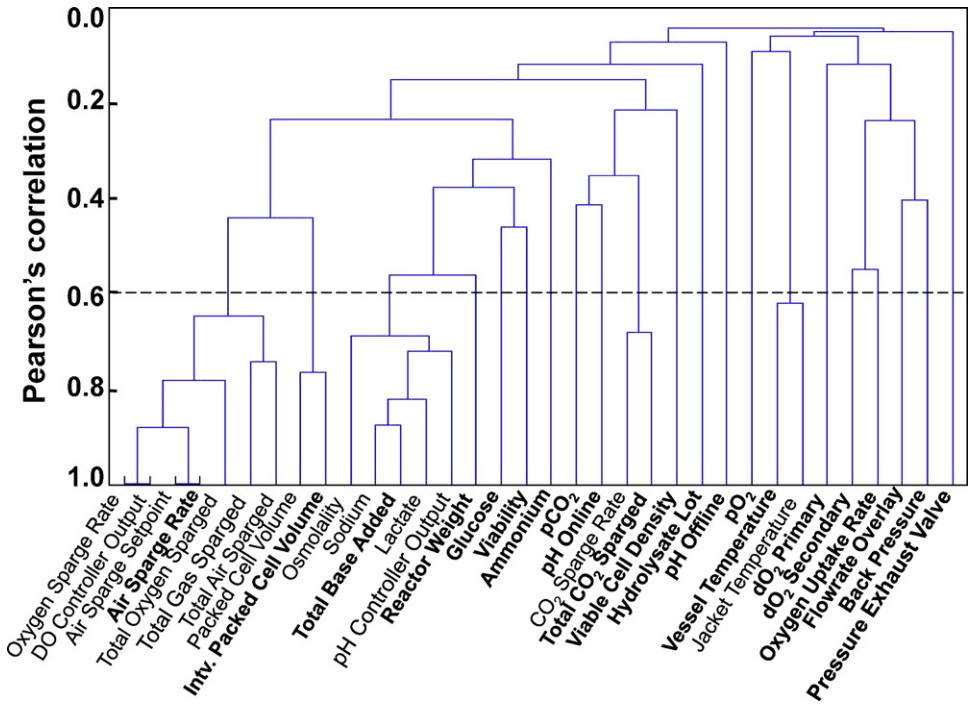

Fig. 7. Dendrogram obtained from hierarchical clustering with weighted average linkage (WPGMA) of process parameters based on their similarity up to day 9 at the 12,000 L scale. Parameters representing each cluster formed at a correlation cut-off value of 0.6 are in bold.

3.5.2. *Association between early stage process parameters and process outcome*

The parameters which are indicative of process outcome in the early stages of a process run can potentially be used as an early warning system for abnormality. The concentrations of lactic acid and sodium ion, total base added, and pH controller output have high weighting factors at the early stages, i.e., day 3 (dataset 4) and day 5 (dataset 5) (Fig. 8). The profiles for lactic acid concentration and total base added in the first five days are shown in Fig. 9b. A greater accumulation of lactic acid is observable in runs with low titer, which correlates with the greater amount of base added as a pH control response. The early differences at the 12,000 L scale between high- and low-titer runs prompted us to investigate the process data from the inoculum train to identify cues for the departures in process outcome. The Spearman's correlation coefficients (*rho*) of process parameters at the 80 L, 400 L, and 2000 L scales were examined. Unlike the observations at the 12,000 L scale, nearly 90% of the parameters at each inoculum scale have *rho* less than 0.1, indicating that these parameters are not strongly correlated to the final titer. An exception is the lactic acid profile at the 2000 L and 400 L scales (Fig. 10a). The concentration of lactic acid is higher in the runs with low final titer. The higher lactic acid concentrations are especially noticeable for the five lowest titer runs. At the 80 L scale, viable cell density and viability also exhibit deviations between high- and low-titer runs (Fig. 10b). These observations are striking in that they suggest that the history of the inoculum may play an important role in determining the final process outcome.

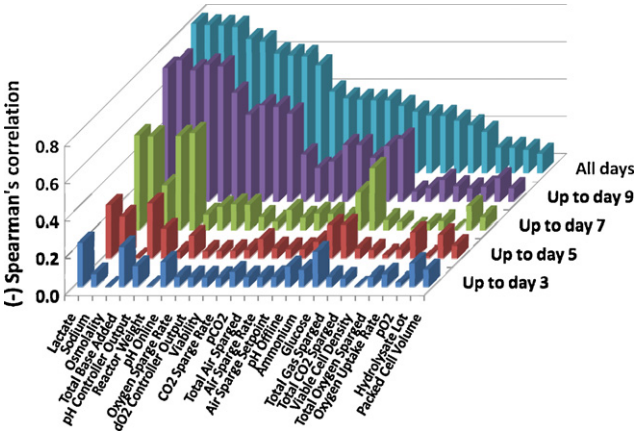

Fig. 8. Relative importance of process parameters at five different stages (up to day 3, day 5, day 7, day 9 and all days) of the production phase (12,000 L). The Spearman's correlation coefficient (*rho*) for relatively important process parameters (with a negative *rho* greater than 0.1 in at least one stage) are shown. For each parameter, *rho* is an average of the ten estimates during 10-fold model cross-validation (see Fig. 3 and Section 2).

4. Discussion

Systematic analysis of large warehouses of manufacturing-scale bioprocess data presents substantial challenges and opportunities to increase process understanding. We describe a framework for systematically interrogating large volumes of process data to identify the hidden characteristics that may be associated with process outcome. A support vector algorithm was employed in this study to construct predictive models for process outcome using parameters measured and archived at the inoculum and production scale bioreactors. Since productivity is a continuous value and not a discrete or binary class, a regression method based on the support vector algorithm was implemented. Support vector machines have been widely used for multivariate data mining due to their strong mathematical foundations, high accuracy and scalability on high-dimensional datasets (Ben-Hur et al., 2008). Further, the kernel-based approach employed here for comparing process runs is readily compatible with the optimization framework of support vector machine. This approach can deal with a large number of parameters without compromising their temporal dynamics. This also allows combination of heterogeneous parameter types such as on-line and off-line temporal parameter profiles, as well as single-point parameters. The notion of constructing models by integrating

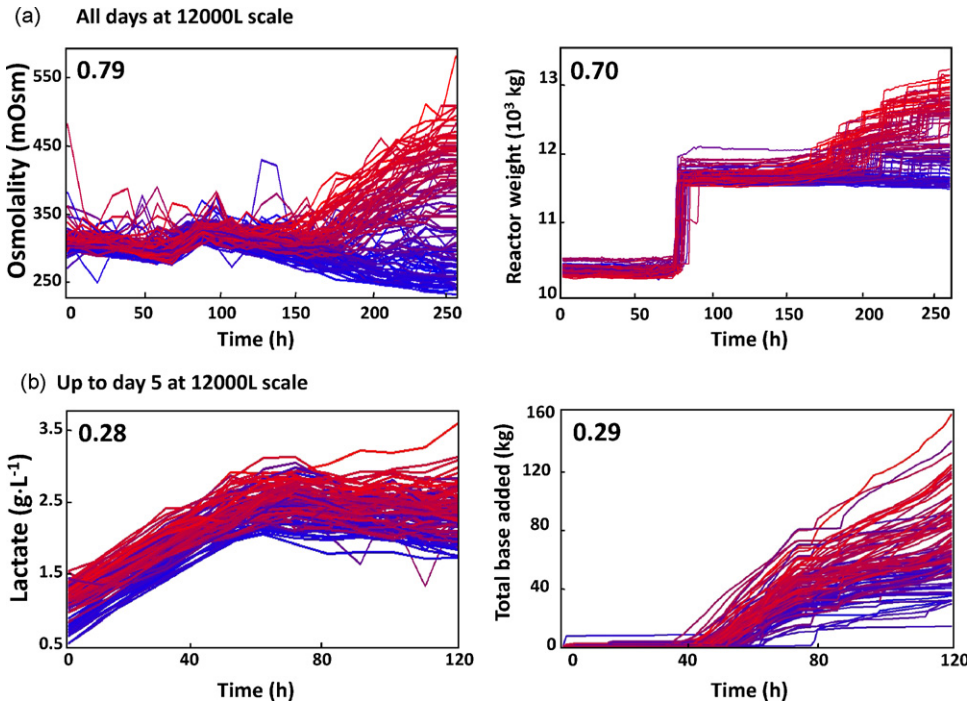

Fig. 9. Selected critical process parameters at different stages of the production scale (12,000 L). (a) Process data for all days. The profiles of osmolality (left panel) and reactor weight (right panel) are shown. (b) Process data up to day 5. The profiles for lactic acid concentration (left panel) and total base added (right panel) are shown. The negative of *rho* for each parameter is displayed in each panel. A red-to-blue gradient is used to label runs in ascending order of the final titer (i.e., runs with low titer are colored in red; runs with higher titer are colored in blue). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

diverse data types has been used in several bioinformatics applications, such as predicting protein–protein interactions (Jansen et al., 2003) and predicting gene functions (Troyanskaya et al., 2003).

In integrating the similarity matrices, it is important that each process parameter is weighted according to its relative contribution in distinguishing the process outcome. Using the Spearman's correlation coefficient, a weighting scheme was employed to assess the outcome predictability of each process parameter. Since the Spearman's rank correlation (*rho*) reflects each parameter's relative contribution to process outcome, it is well-suited for determi-

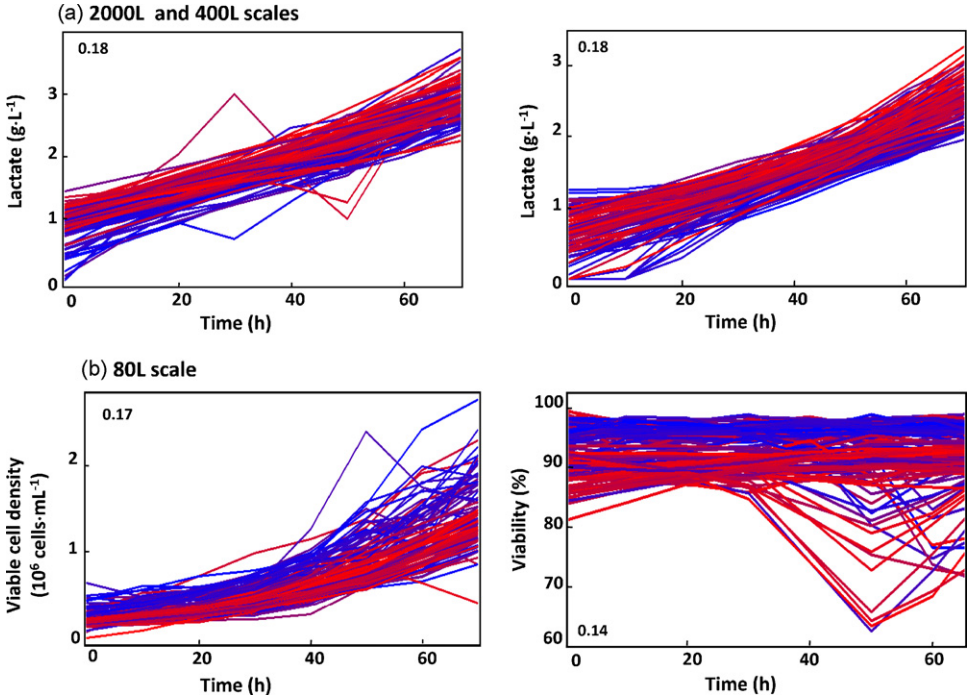

Fig. 10. Critical process parameters measured at the inoculum scales (80 L, 400 L, and 2000 L). (a) The profiles of lactic acid concentration at the 2000 L (left panel) and the 400 L (right panel) scales are shown. (b) The profiles at the 80 L scale of viable cell density (left panel) and viability (right panel) are shown. The negative of *rho* for each parameter is displayed in each panel. A red-to-blue gradient is used to label runs in ascending order of the final titer. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of the article.)

nation of critical process parameters whose atypical behavior is likely to influence process outcome. These critical parameters can provide additional criteria for evaluating the overall performance of a process run in addition to the prescribed process outcome, in our case, the end-point titer. In this study, process data was organized in a sequential manner, which allows us to determine critical parameters at different stages of a run. Prominent among the parameters at the inoculum stage are lactate, viability, and viable cell density, all of which have *rho* > 0.1 at the 400 L scale. At the 12,000 L scale, lactic acid and base addition profiles as early as at day 3 are indicative of the final titer. Also, parameters such as pH, osmolality, DO controller output, viability, and reactor weight have high *rho* at late stages of the production bioreactor. Many of these critical parameters are mutually correlated as identified in this study. Dimensionality reduction was achieved by abbreviating the correlated parameters by a single representative parameter. Models constructed using a smaller subset of relatively independent parameters resulted in high predictabilities, which were comparable to the all-parameter models. This further suggests that the dominant correlations within process parameters can be highlighted by a smaller subset of parameters. Among the correlated parameters, lactic acid, sodium, osmolality, pH controller output and base addition profiles are significantly correlated at late stages of production, highlighting that higher lactate production and consequently higher base consumption, osmolality and lower viabilities are observed during the late stages of runs with lower final titer.

The adverse effects of lactic acid accumulation on viability and recombinant protein productivity of mammalian cells are well-known. Accumulation of higher levels of lactic acid is an impediment to achieving high cell densities, and therefore, high product titers. Thus, lactic acid concentration being one of the critical factors is not surprising. What was unexpected was that differences between the lactic acid profiles of high and low-titer runs emerge at very early stages of a train of cultures in different reactors. Employing process data of cell expansion stage in the inoculum train and the first three days of the final production stage, our model predicts process outcome with good accuracy (Fig. 5). Even employing only inoculum train data, the outcome can be predicted, albeit with lower accuracy. This finding strongly suggests that key events affecting process outcome occur before day 3 in the production-scale bioreactor.

Critical process parameters identified in this study are correlated to productivity. However, the underlying causes of these correlations are unclear. Different lactate levels at early stages of culture may be a result of differences in inoculated cells from inoculum train, or different culture conditions (i.e., parameter profiles) at the early stage after inoculation. The former implies that the history of cells exerts an effect on metabolism which is reflected in lactate production and eventually the productivity. The later suggests that the profile of some parameter(s) causes metabolic perturbations, resulting in higher lactate production. The causal parameter may be, or may not be, one of those already measured. A closer scrutiny of the early stage data may provide further clues on the potential causes of the deviations in cell culture performance. Furthermore, incorporation of additional data, either by increasing the number of process parameters examined for each run, or increasing the total number of runs in the dataset can increase model predictability, especially at early stages of the run in the inoculum train. A preliminary study performed based on the on-line and off-line process data from a subset of 30 runs was unable to predict cell culture performance at early stages of the inoculum train with good accuracy. However, for the same dataset of 30 runs, the prediction accuracy was high when data from all the stages of the run (including inoculum train and production) was included.

The potential parameters causing process outcome variations should be experimentally verified. Carrying out such experiments in manufacturing scale reactors is extremely costly, let alone experimenting on the entire production train. Nevertheless the result of this analysis can provide crucial information for further investigation.

5. Concluding remarks

Many life-saving therapeutics today are commercially produced in state-of-the-art manufacturing facilities with automated control systems for measuring and archiving a plethora of process parameters. The historical archives of these datasets present vast data mining opportunities to unearth and better understand the hidden correlations between process inputs and outputs such as product quality and yields. This study describes a multivariate data mining technique that combines more than one-hundred time-dependent off-line, on-line, as well as single point parameters across different production stages to predict a key process output—the run productivity. Overall, this study demonstrates the power of mining process data in revealing hidden correlations between process outcome and process parameters. The generated insights will certainly lead to hypotheses for further investigations and potentially lead to intervention strategies to render the process more robust.

Acknowledgement

The computational resources for this work were provided Minnesota Supercomputing Institute at University of Minnesota, Twin Cities.

References

Aggarwal, S., 2009. What's fueling the biotech engine—2008. Nat. Biotech. 27 (11), 987–993.
Ambroise, C., McLachlan, G.J., 2002. Selection bias in gene extraction on the basis of microarray gene-expression data. Proc. Natl. Acad. Sci. U.S.A. 99 (10), 6562–6566.
Bachinger, T., Riese, U., Eriksson, R., Mandenius, C.F., 2000a. Monitoring cellular state transitions in a production-scale CHO-cell process using an electronic nose. J. Biotechnol. 76 (1), 61–71.
Bachinger, T., Riese, U., Eriksson, R.K., Mandenius, C.F., 2000b. Electronic nose for estimation of product concentration in mammalian cell cultivation. Bioprocess Biosyst. Eng. 23 (6), 637–642.
Bakshi, B.R., Stephanopoulos, G., 1994. Representation of process trends. 4. Induction of real-time patterns from operating data for diagnosis and supervisory control. Comput. Chem. Eng. 18 (4), 303–332.
Ben-Hur, A., Ong, C.S., Sonnenburg, S., Scholkopf, B., Ratsch, G., 2008. Support vector machines and kernels for computational biology. PLoS Comput. Biol. 4 (10), e1000173.
Beyer, K.S., Goldstein, J., Ramakrishnan, R., Shaft, U., 1999. When is "nearest neighbor" meaningful? Proceedings of the 7th International Conference on Database Theory, pp. 217–235.
Chang, C.C., Lin, C.J., 2001. LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
Charaniya, S., Hu, W.S., Karypis, G., 2008. Mining bioprocess data: opportunities and challenges. Trends Biotechnol. 26 (12), 690–699.
Coleman, M.C., Block, D.E., 2006. Retrospective optimization of time-dependent fermentation control strategies using time-independent historical data. Biotechnol. Bioeng. 95 (3), 412–423.
Coleman, M.C., Buck, K.K., Block, D.E., 2003. An integrated approach to optimization of *Escherichia coli* fermentations using historical data. Biotechnol. Bioeng. 84 (3), 274–285.
Glassey, J., Montague, G.A., Ward, A.C., Kara, B.V., 1994a. Artificial neural network based experimental design procedures for enhancing fermentation development. Biotechnol. Bioeng. 44 (4), 397–405.
Glassey, J., Montague, G.A., Ward, A.C., Kara, B.V., 1994b. Enhanced supervision of recombinant *E. coli* fermentations via artificial neural networks. Process Biochem. 29 (5), 387–398.
Huang, J., Nanami, H., Kanda, A., Shimizu, H., Shioya, S., 2002. Classification of fermentation performance by multivariate analysis based on mean hypothesis testing. J. Biosci. Bioeng. 94 (3), 251–257.
Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., Gerstein, M., 2003. A Bayesian networks approach for predicting protein–protein interactions from genomic data. Science 302 (5644), 449–453.

Kamimura, R.T., Bicciato, S., Shimizu, H., Alford, J., Stephanopoulos, G., 2000. Mining of biological data. II. Assessing data structure and class homogeneity by cluster analysis. Metab. Eng. 2 (3), 228–238.

Kirdar, A.O., Conner, J.S., Baclaski, J., Rathore, A.S., 2007. Application of multivariate analysis toward biotech processes: case study of a cell-culture unit operation. Biotechnol. Prog. 23 (1), 61–67.

Muller, K.R., Mika, S., Ratsch, G., Tsuda, K., Scholkopf, B., 2001. An introduction to kernel-based learning algorithms. IEEE Trans. Neural Networks 12 (2), 181–201.

Scholkopf, B., Smola, A.J., Williamson, R.C., Bartlett, P.L., 2000. New Support Vector Algorithms. MIT Press, pp. 1207–1245.

Stephanopoulos, G., Locher, G., Duff, M.J., Kamimura, R., Stephanopoulos, G., 1997. Fermentation database mining by pattern recognition. Biotechnol. Bioeng. 53 (5), 443–452.

Troyanskaya, O.G., Dolinski, K., Owen, A.B., Altman, R.B., Botstein, D., 2003. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in *Saccharomyces cerevisiae*). Proc. Natl. Acad. Sci. U.S.A. 100 (14), 8348–8353.

Usama, F., Ramasamy, U., 1996. Data mining and knowledge discovery in databases. Commun. ACM 39 (11), 24–26.

Vapnik, V.N., 1998a. Statistical Learning Theory. Wiley-Interscience, New York, 736 pp.

Vapnik, V.N., 1998b. Statistical Learning Theory [M]. New York.

Vlassides, S., Ferrier, J.G., Block, D.E., 2001. Using historical data for bioprocess optimization: modeling wine characteristics using artificial neural networks and archived process information. Biotechnol. Bioeng. 73 (1), 55–68.