SvedbergOpen
DISSEMINATION OF KNOWLEDGE
International Journal of Artificial Intelligence and Machine Learning
Publisher's Home Page: https://www.svedbergopen.com/
Research Paper Open Access

Data-Driven Control Strategies In Upstream Bioprocessing: Integrating Analytics, Automation, And Adaptive Systems

Sai Surya Raja Amogha Tenneti

Eurofins PSS, USA

Abstract
Driven by process analytical technology (PAT), advanced sensor instrumentation, and multivariate statistical modeling, data-driven control strategies have opened up the next generation of biological manufacturing, enabling unprecedented levels of real-time process observation, comprehension, and governance in upstream bioprocessing through continuous, high-resolution perception into key process parameters and quality attributes. Control architectures, ranging from basic feedback control to complex model predictive control, translate these analytics-driven results into coordinated and timed actions for the process equipment to balance the highly nonlinear and multivariable nature of cell culture processes. Soft sensors, anomaly detection, and predictive quality models, as well as adaptive control policies, have been developed from historical and real-time process data using machine learning-based approaches. The advent of digital twin software for biomanufacturing will lead to mechanistic models and data analysis being integrated into dynamic digital twins that can be optimized and, subsequently, scaled up and deployed in real-time in response to any deviations throughout the lifecycle of the bioprocess. Therefore, these platforms are gradually leading biomanufacturing towards more adaptive and autonomous biomanufacturing environments with higher process consistency, product quality, and operational efficiency. Issues with data integrity, model life cycle governance, workforce capability development, and regulatory acceptance of automated decision-making systems must be addressed if the potential is to be realized on a commercially viable basis. The continual maturation of integrated data-driven control architectures will become the norm for the biologics manufacturing of the future, driving new standards for process performance and quality assurance.

Keywords: Process Analytical Technology, Model Predictive Control, Digital Twin, Machine Learning, Upstream Bioprocessing

1. Introduction

Overall, upstream bioprocessing has undergone a model shift over the last 20 years, predominantly driven by advances in analytical technologies, computational modeling, and automation systems. Biological manufacturing processes are complex systems that involve living cellular entities that adapt dynamically to environmental and nutritional perturbations and are affected by intrinsic biological variability. Historically, bioreactor process control research has focused on periodic off-line sampling, heuristic rules, and operator experience, which may not be sufficient to ensure consistent product quality and process reproducibility [1], [2].

Retrospective monitoring shortcomings became clearer with the increasing manufacturing of products such as monoclonal antibodies, recombinant proteins, and cell and gene therapy products, which all require tighter process control, increased process understanding, and more stringent quality assurance. Regulatory guidelines and requirements evolved in parallel toward a more process analytical technology (PAT) and quality by design (QbD)-focused approach to modern biopharmaceutical development [3].

Such data-driven control systems are in contrast to those that use historical data and analyze it, determining which variables to modify in a process and possibly doing so before they are outside of a given range. In particular, such systems may allow for simultaneous optimization of process variables. The integration of multivariate data analysis and machine learning with digital twin technology and advanced control algorithms has given rise to a new class of adaptive, even autonomous bioprocessing systems [4] [5].

This review will discuss a systematic approach toward data-driven control of upstream bioprocessing, with the aim of providing a thorough overview of the analytics, control theory, and computational and digital technologies that enable such programs. The review will also highlight the challenges in the process of implementation and pathways to fully automated biomanufacturing. Each section will summarize the relevant literature, providing a technical and practical overview of such a rapidly evolving area.

2. Analytical Foundations and Process Analytical Technology

The successful implementation of data-driven control strategies relies on the continuous availability of high-quality data. According to regulatory definitions, Process Analytical Technology means the use of online or in-line analytical instruments in order to foster understanding of manufacturing processes and to ensure product quality proactively rather than retrospectively [3], [6]. The move from end-product testing to in-process monitoring has also fueled the use of sensors and inline analytical instrumentation in manufacturing operations.

Modern bioreactor systems routinely include sensors that can measure many process parameters and critical quality attributes, with a long history of accurate measurement of physical parameters such as temperature, pH, and dissolved oxygen by strong electrochemical probes. Recently, in online process control, the measurement of the biomass concentration by capacitance, the dissolved carbon dioxide by infrared spectroscopy, and the important metabolites glucose, lactate, glutamine, and glutamate by near-infrared and Raman spectroscopic measurements have been added to the existing process analyzers [7], [8]. Spectroscopy-based approaches are non-intrusive and continuous or quasi-continuous in nature and allow simultaneous acquisition of several analytes.

But the abundance of available data is also a burden, as multi-parameter, high-frequency data needs to be appropriately handled, processed, and interpreted to convert them into useful and relevant information. Raw spectroscopic data are subject to noise, baseline drift, and matrix effects due to the presence of matrix components in the measurement, thus requiring chemometric pretreatment and the development of calibration models [9]. Hence, the success of process control decisions depends upon the quality of the analytical measurements on which they are based.

Multivariate statistical procedures have often been used for providing information to either the operator or the automatic control system about the process under study. For example, principal component analysis (PCA) is often used for dimensionality reduction and visualization of the process trajectory in latent variable space to detect deviations from normal operating conditions [10]. Partial least squares (PLS) regression extends this idea to model-based prediction of process variables or quality attributes that cannot be measured directly and are predicted from surrogate analytical signals. Such chemometric tools are now ubiquitous in PAT implementations in industrial bioprocessing applications. Real-time data acquisition and these chemometric models are the analytical backbone of data-derived control strategies.

Table 1: Key Analytical Technologies in Real-Time Bioprocess Monitoring [6, 7]

3. Control Theory Frameworks: From PID to Model Predictive Control

With the basic analysis in place, the control theory tools for transforming the information obtained in real time into bioprocess control can now be reviewed. The predominant industrial bioprocess control still uses the proportional-integral-derivative (PID) controller today, a classical feedback controller which considers the amplitude, the integral (accumulated offset), and the derivative (rate of error change) of the error in

manipulating the control variable [5]. Conversely, setpoint-maintaining control loops for parameters such as pH and dissolved oxygen are amenable to a PID controller since each is a relatively well-understood single-input, single-output system.

However, PID control may not be an appropriate solution, as bioprocesses are characterized by their nonlinear dynamics, their time-varying nature, and strong interaction between process variables. For instance, a change in feeding rate affects not only the concentration of the nutrient in the feed vessel but also the growth rate of the cells, build-up of metabolic by-products, oxygen requirement, and even such product quality attributes as glycosylation [5], [6]. Because a PID controller typically tackles a single variable at a time, this method is not suitable.

Model Predictive Control (MPC) is based on a different control philosophy. Rather than attempting to control the process based on perturbations in the plant variables, the MPC uses a model of the process to predict its behavior over a prediction horizon. At each control step, the MPC will solve an optimization problem that determines the sequence of control inputs that minimizes a user-defined objective function subject to operating constraints [11]. This modeling ability allows the controller to predict the results of control actions in real-time, which is useful for multivariable systems.

Model types used in MPC frameworks include mechanistic models, which are first empirical models that are data-driven, and hybrid models, which make use of a mechanistic model framework but with empirical model parameters. There are trade-offs associated with the predictive accuracy, the computational cost, and the ease of identification and maintenance of these models [12]. In bioprocessing applications, hybrid models, where known stoichiometric and kinetic behavior is complemented with data-based parts in order to account for unmodeled dynamics, have been shown to be particularly promising in improving generalization across operating conditions.

Some adaptive MPC formulations extend standard implementations by allowing process models to be updated online as process data become available. In bioprocessing applications, for example, the dynamics of a cell culture process may differ substantially for different cell lines or media formulations or for different scales of operation. By continuously updating the internal model to track the change in behavior of the underlying process, adaptive MPC systems can maintain their predictive and control performance over the entire operating range of a culture run and across different campaigns [11], [12]. Hence, MPC with state-of-the-art real-time analysis provides a coherent and powerful model for predictive and adaptive upstream process control.

Table 2: Comparison of PID and Model Predictive Control in Bioprocessing [5, 6]

4. Machine Learning Applications in Bioprocess Optimization

Machine learning applications in upstream bioprocessing have grown considerably in recent years due to the increasing availability of historical data and computational power. Machine learning can outperform simpler methods when data is high-dimensional, nonlinear effects are present, or there are patterns that cannot easily be formulated in a mechanistic model [6], [13].

Soft sensors and mass- or energy-balance models have been implemented using supervised learning algorithms to estimate process variables that are difficult or costly to measure. Neural networks, support vector regression, and ensemble methods such as random forests have been trained using historical process data to predict cell viability, product titer, or critical quality attributes from combinations of readily available process measurements [13]. After validation, these models can be used as real-time estimates in the control loop to increase the analytical capability of the monitoring system without needing additional measurements.

Unsupervised learning methods such as clustering algorithms or other dimensionality reduction methods, including more complex versions of PCA, can be used for process characterization and anomaly detection. By learning the statistical structure of normal process behavior from historical data, these methods are able to detect anomalies or abnormal process conditions early on and trigger alerts or corrective actions before any important negative impact on quality or yield can occur [10]. The temporal nature of bioprocess data has also led to the application of recurrent neural networks and long short-term memory architectures, which are able to capture temporal dependencies, hence making the models amenable to modeling operations occurring during cell culture [13].

Reinforcement learning is a very different setting of machine learning in which the agent learns to control a process by interacting with its environment. Set up as a reinforcement learning task, the agent learns to select actions such as adjusting the feed rate or changing the setpoint of the process temperature based on the reward signal received from the observations of the process [6]. Unlike supervised learning that typically relies on labeled training data that reflect the desired operating policy, reinforcement learning allows for learning new non-intuitive control strategies through exploration of the action space and deduction of action preferences with respect to the process. The application of reinforcement learning on a live bioprocess is limited by the challenge of exploring the production environment, and training through simulation software is used to develop and validate policies before implementation [14].

A particularly fruitful direction for further research has been the area of physics-informed machine learning techniques, where mechanistic intuitions are incorporated into the architecture of a machine learning model to avoid needing large amounts of training data and improve extrapolation performance out of the training domain. Hybrid modeling frameworks that couple mechanistic metabolic models with data-driven correction terms have been shown to improve prediction accuracy and robustness over any of the individual approaches [12]. These developments reflect a broader objective of a first-principles understanding being combined with data-driven inference, rather than competing, for bioprocess modeling and control.

Table 3: Machine Learning Methods and Their Bioprocessing Applications

Machine Learning Category	Representative Methods	Bioprocessing Application	Key Benefit
Supervised Learning	Artificial Neural Networks, Random Forest, SVR	Soft sensing, titer prediction, quality attribute modeling	Predictive accuracy from historical data
Unsupervised Learning	PCA, Clustering, Autoencoders	Anomaly detection, process characterization	No labeled data required
Recurrent/Temporal Models	LSTM, Recurrent Neural Networks	Sequential process dynamics modeling	Captures time-dependent behavior
Reinforcement Learning	Q-Learning, Policy Gradient Methods	Adaptive control policy development	Learns from process interaction
Hybrid Modeling	Physics-Informed Neural Networks	Mechanistic-data integration	Improved generalization and interpretability

5. Digital Twin Integration and Autonomous Biomanufacturing

The digital twin has emerged as a common framework to integrate RTDA, mechanistic modeling, and ML in the bioprocess domain. A digital twin is a dynamic virtual representation of a physical process system, which is generally continuously updated through the exchange of data with the real process and capable of running simulations, scenario assessments, and optimizations that would otherwise be prohibitively expensive to operate on the physical system [6], [14]. Digital twins commonly comprise three different modeling approaches in upstream bioprocessing: mass balance models; mechanistic models, which describe kinetic growth and metabolism and thermodynamic and transport relationships; and data-driven models, which capture the residual dynamics not represented in the mechanistic model.

Digital twins can add value throughout the bioprocess development and manufacturing lifecycle. For example, in the process development stage, digital twin platforms can enable in silico navigation through the design

space, enabling the minimization of experimental work required to identify optimal operating points and strong control strategies [14]. This could save time and resources required, which is particularly important in cell culture, as many of the consumables and materials are expensive. Scale-up is another key area of variation and risk, typically addressed empirically. However, digital twin-assisted engineering could use hydrodynamic and mass transfer models to increase understanding of how transfer between scales impacts system behavior and allows for a model-based translation of control strategies [9].

Digital twins are also widely used in industrial applications, e.g., in advanced process control. These digital twins (e.g., chemical process digital twins) maintain estimates of their internal states while also providing estimates of future process trajectories that might be needed for potential control actions (to keep the process in the optimal operating space) [6]. A digital twin can automatically correct the process inputs or issue diagnostics or advisories to operators if there is a deviation from normal operating parameters, depending on the type of deviation and its severity. This is an improvement to customary alarms, which are generally reactive in nature and generate many nuisance alarms that can overwhelm operational decision-making [3].

The end goal of autonomous biomanufacturing is to monitor, analyze, decide, and actuate a closed-loop biomanufacturing process with minimal human intervention in continuous operation. In addition to the technical systems that form the focus of this article, validation of laboratory and process performance, failure mode analysis, and regulatory acceptance frameworks will be key to achieving this goal [2], [6]. Regulators are developing guidance documents on how these digital and advanced process control technologies fit into existing quality systems [3]. There is some good practice guidance around data integrity and cybersecurity, model governance, human oversight, etc. Full automation with no human oversight in a regulated environment poses a bigger challenge.

Table 4: Digital Twin Capabilities Across the Bioprocess Lifecycle [3, 14]

Conclusion

Data-driven control systems have thus transformed upstream bioprocessing from a process governed by empirical convention and intermittent sampling to a process empowered by real-time intelligence, prediction, and optimization. A few analytical examples include inline spectroscopic platforms, capacitance-based biomass estimation systems, and chemometric process data pipelines, which leverage continuous high-quality process information to enable new control systems that characterize real-time bioprocessing. Methods based on classical feedback control theory and model predictive control (MPC) have realized systems and process interventions that respect coordination and constraints and maintain biological systems in their optimal operating regimes under different culture conditions. Advanced machine learning approaches are also able to extract relevant predictive information from high-dimensional process data. Hybrid modeling strategies that capitalize on mechanistic knowledge of biological systems and data-driven approaches provide increased predictive performance, robustness, and interpretability. Digital twin platforms represent a computational ecosystem that combines analytics, control theory, and machine learning to enable scenario-based optimization, anticipation of deviations, and holistic scale-up mapping throughout bioprocess development and manufacturing. While autonomous biomanufacturing represents a desirable end-state, the advancing enabling conditions of data governance, model validation approaches, cybersecurity architecture, and regulatory model evolution will contribute to the realization of autonomous decision-making in quality- and compliance-sensitive manufacturing. In parallel with these opportunities, data-driven and model-based control strategies will set the future standard for biologics manufacturing, enabling quantitative improvements to process consistency, product quality, development efficiency, and the ability of the manufacturing industry to meet therapeutic demand at commercial manufacturing scale.

References

[1] Eric S. Langer and Ronald A. Rader, "Continuous Bioprocessing and Perfusion: Wider Adoption Coming as Bioprocessing Matures," BioProcess International, 2014. [Online]. Available: https://bioprocessingjournal.com/continuous-bioprocessing-and-perfusion-wider-adoption-coming-as-bioprocessing-matures/#:~:text=Among%20the%20technologies%20now%20gaining,technology%2C%20in%20terms%20of%20adoption.

[2] International Conference on Harmonisation, "ICH Harmonised Tripartite Guideline Pharmaceutical Development Q8(R2)," International Conference On Harmonisation Of Technical Requirements For Registration Of Pharmaceuticals For Human Use, 2009. [Online]. Available: https://www.ich.org/page/quality-guidelines
[3] U.S. Food and Drug Administration, "Guidance for Industry: PAT — A Framework for Innovative Pharmaceutical Development, Manufacturing, and Quality Assurance," FDA, Rockville, MD, 2004. [Online]. Available: https://www.fda.gov/media/71012/download
[4] A. S. Rathore et al., "Process analytical technology (PAT) for biopharmaceutical products," Analytical and Bioanalytical Chemistry, 2010. [Online]. Available: https://doi.org/10.1007/s00216-010-3781-x
[5] Jarka Glassey et al., "Process analytical technology (PAT) for biopharmaceuticals," Biotechnology Journal, 2011. [Online]. Available: https://doi.org/10.1002/biot.201000356
[6] Xiaoyan Guo et al., "Stepwise increase in the production of 13R-manoyl oxide through metabolic engineering of Saccharomyces cerevisiae," Biochemical Engineering Journal, 2019. [Online]. Available: https://doi.org/10.1016/j.bej.2019.01.012
[7] Xiaoli Xu et al., "Integration of electrochemistry in micro-total analysis systems for biochemical assays: Recent developments," Talanta, 2009. [Online]. Available: https://doi.org/10.1016/j.talanta.2009.06.039
[8] Patrick Wechselberger et al., "Real-time estimation of biomass and specific growth rate in physiologically variable recombinant fed-batch processes," Bioprocess and Biosystems Engineering, 2013. [Online]. Available: https://doi.org/10.1007/s00449-012-0848-4
[9] Carl-Fredrik Mandenius and Anders Brundin, "Bioprocess optimization using design-of-experiments methodology," Biotechnology Progress, 2008. [Online]. Available: https://doi.org/10.1002/btpr.67
[10] Paul Nomikos and John F. MacGregor, "Monitoring batch processes using multiway principal component analysis," AIChE Journal, 1994. [Online]. Available: https://doi.org/10.1002/aic.690400809
[11] S. Joe Qin and Thomas A. Badgwell, "A survey of industrial model predictive control technology," Control Engineering Practice, 2003. [Online]. Available: https://doi.org/10.1016/S0967-0661(02)00186-7
[12] Kai Shen and Joseph K. Scott, "Rapid and accurate reachability analysis for nonlinear dynamic systems by exploiting model redundancy," Computers & Chemical Engineering, 2017. [Online]. Available: https://doi.org/10.1016/j.compchemeng.2017.08.001
[13] Laura M. Helleckes, et al., "Machine learning in bioprocess development: from promise to practice." Trends in Biotechnology, 2023. Available: https://www.sciencedirect.com/science/article/abs/pii/S0167779922002815
[14] Sara Sousa Rosa et al., "Maximizing mRNA vaccine production with Bayesian optimization," Biotechnology and Bioengineering, 2022. [Online]. Available: https://doi.org/10.1002/bit.28216
[15] Anurag S Rathore and Helen Winkle, "Quality by design for biopharmaceuticals," Nature Biotechnology, 2009. [Online]. Available: https://doi.org/10.1038/nbt0109-26