Phenotype analysis of cultivation processes via unsupervised learning: demonstration for *Clostridium pasteurianum*

Yaeseong Hong, Tom Nguyen and An-Ping Zeng

TUHH Hamburg University of Technology

BIO PROCESS AND SYSTEMS ENGINEERING TUHH

Introduction

“What is the typical cultivation behavior of your strain?”

The answer may be simple for strains with robust fermentation behavior with little dynamic variations. But what if the strain shows heterogeneous phenotypes that you need to pinpoint from multiple cultivation experiments? In this work, a clustering method of fermentation data is presented that analyses multiple cultivation datasets based on the two hypotheses: first, stationary cellular phenotypic behavior can be described as a vector consisting of specific rates; second, similar vectors can be clustered as a representative phenotypic group surrounding a centroid.

As a field of unsupervised learning, clustering can be utilized as a potent tool that can pinpoint different types of phenotypic behavior as clusters. Conditional triggers of specific phenotypes as well as differentiation of cellular behaviors can be outlined. For its demonstration, fermentation data of *Clostridium pasteurianum* strains were used.

Material and Methods

The used fermentation data were generated from 44 small scale cultivations (≥ 50 mL) and 46 controlled fermentation processes (≤ 3 L) of different *C. pasteurianum* strains. MATLAB 2020b was used for the calculative process (Fig. 1). Briefly, cultivation data were used to interpolate inter-sample data points via Piecewise Cubic Hermite Interpolating Polynomial (PCHIP). After estimation of the time-derivatives of concentration data, specific consumption or production rates were calculated. Followed by outlier removal (1.6 percentiles), the number of clusters were estimated using the silhouette evaluation with cosine distance metric and z-score normalization. The resulting clusters were calculated as medians from clustered data.

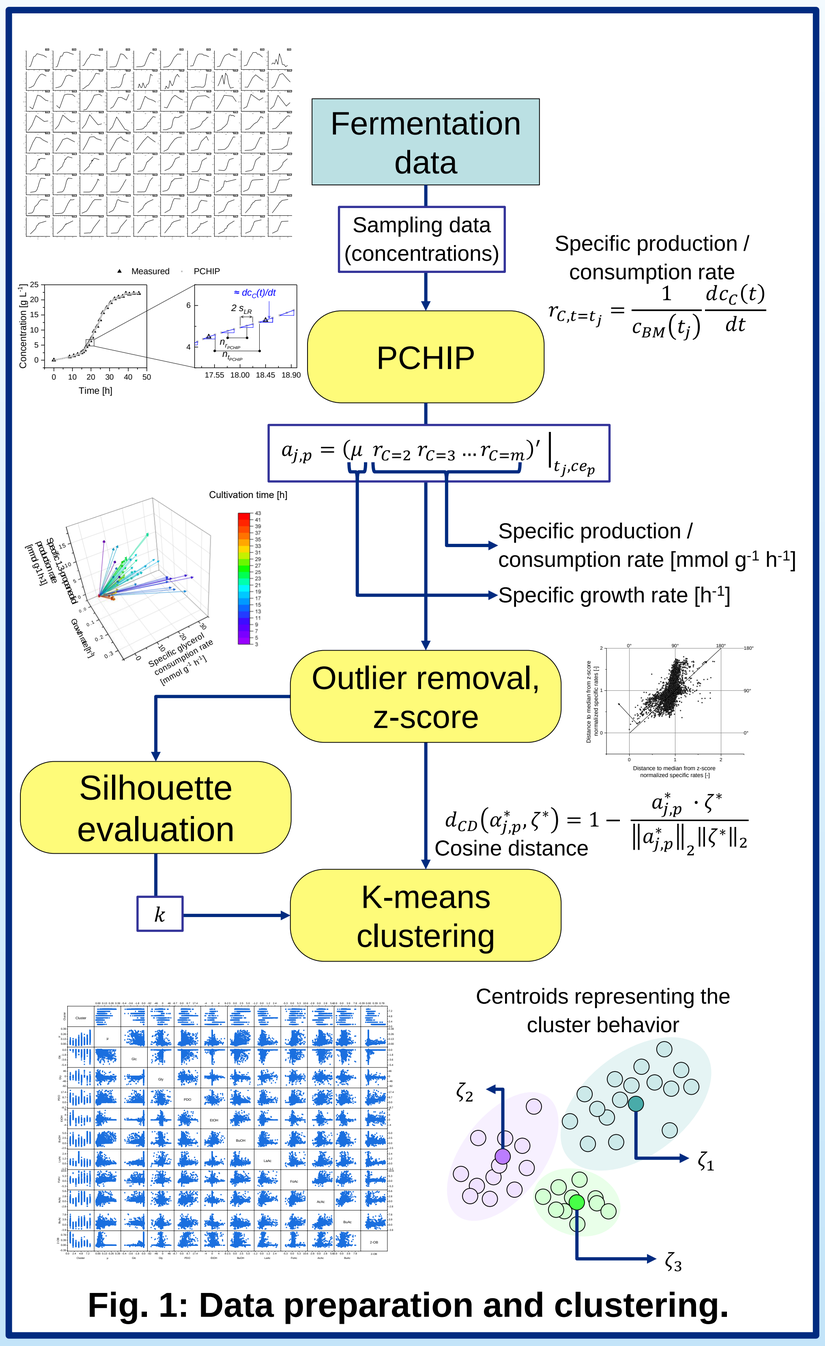

Fig. 1: Data preparation and clustering.

Results

For the demonstrated case, **14 clusters** as characteristic phenotypes were found (Fig. 2). Analysis of conditionality of phenotypic manifestations showed potential trends (e.g. initial $FeSO_4 \cdot 7H_2O$ concentration) or unique factors (e.g. distribution of strains for a specific cluster).

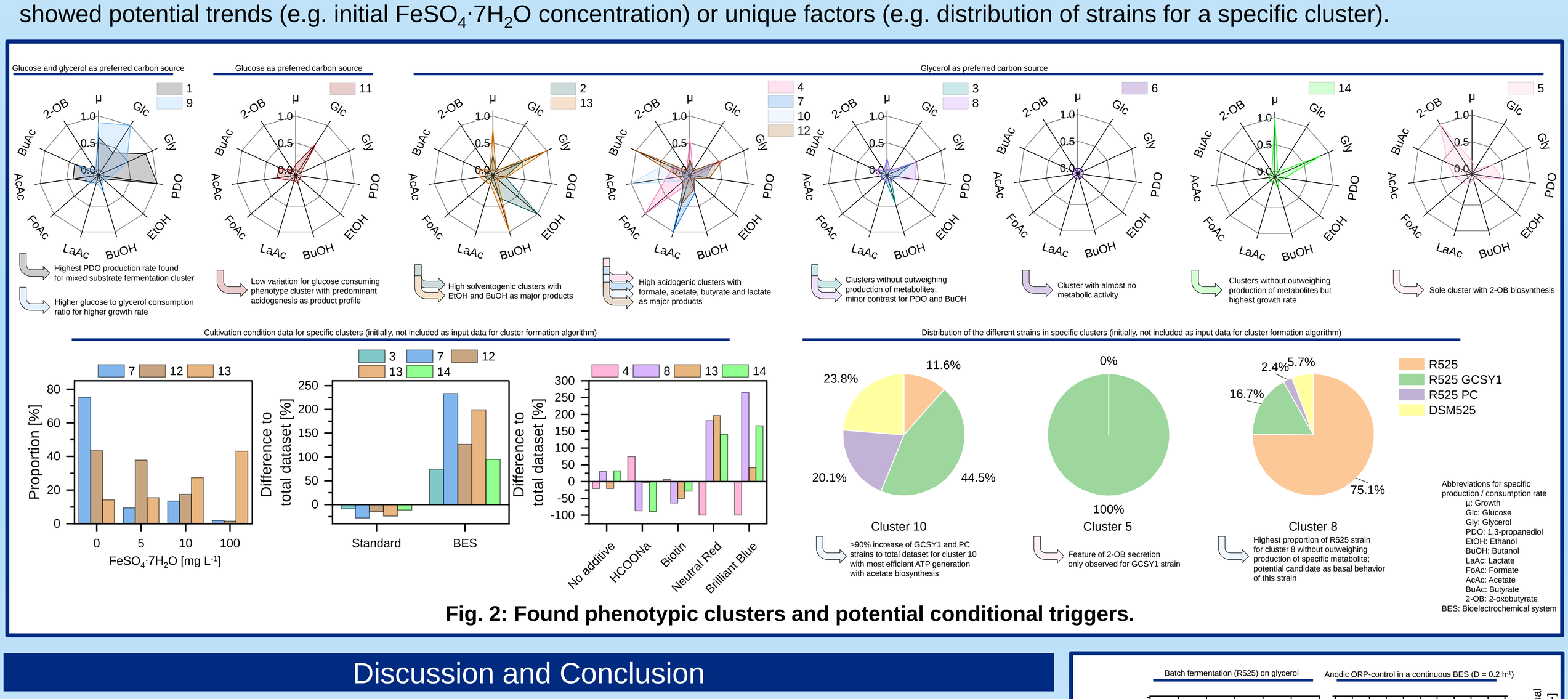

Fig. 2: Found phenotypic clusters and potential conditional triggers.

Discussion and Conclusion

High dimensionality of the datasets coupled to heterogeneity of vector densities leads to unsatisfactory results using density-based clustering (DBSCAN). Additional data treatment leads to ‘weighted’ calculation of distances for k-means clustering.

Characteristic phenotypical behavior in 14 different clusters were captured. However, data treatment and clustering parameter optimization are required for distinction between characteristic behavior and data noise. Potential cause and effect relations can be analyzed via inclusion of additional information that was not used for clustering.

Besides from analysis for conditional triggers, superposition of the identified clusters can be used to analyze cultivation process (here calculated as nonnegative linear least-squares problem). Dynamic shifts of predominant clusters or shifts of superposition-based cluster distribution for different steady state conditions (e.g. ORP control via anodic BES) can be used as additional tool to capture minor differences (Fig. 3). Further tuning and optimization is currently in progress.

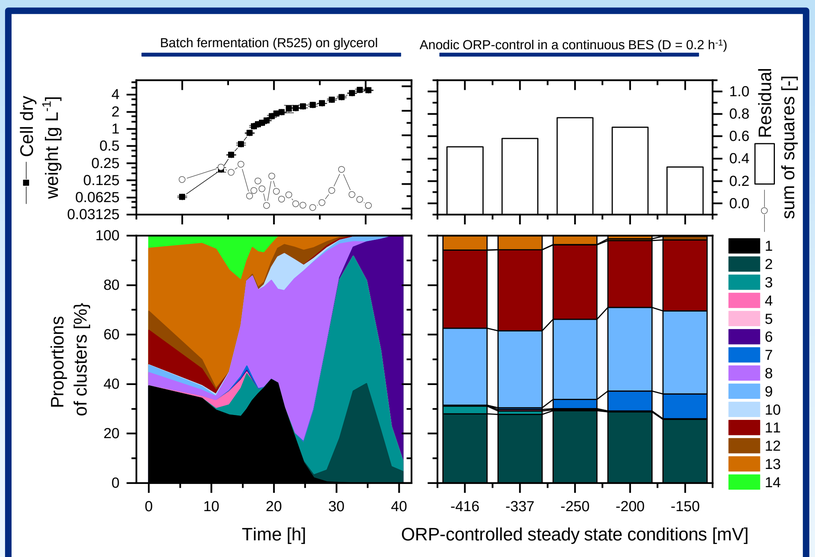

Fig. 3: Superposition-based analysis.

Contact

	Yaeseong Hong,	Prof. An-Ping Zeng
Email:	yaeseong.hong@tuhh.de	AZE@tuhh.de
Phone:	+49 (0)40 428 78 4401	- 4183

Institute of Bioprocess and Biosystems Engineering,
Hamburg University of Technology, Denickestrasse 15, 21073 Hamburg, Germany