Data-Based Clustering and Control of Similar Biological Systems

Peilin Zhang , Antonis Papachristodoulou, Idris Kempf

University of Oxford

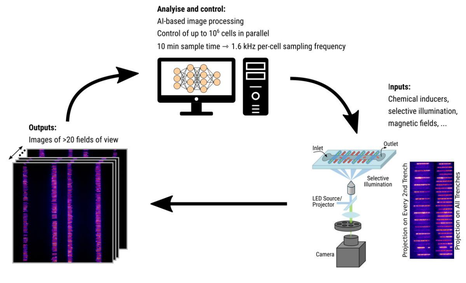
1. Motivation

Cybergenetic gene expression control enables applications in engineering biology, drug development, and biomanufacturing. Microfluidic devices enable control of millions of cells in parallel, but throughput requirements and biological system heterogeneity limit algorithm choice in practice.

In this study, we develop a similarity-aware control framework to increase computational efficiency of cybergenetic control applications, and implement data-based control to accommodate biological system heterogeneity. We cluster cells according to similarity and develop a hierarchical leader-follower architecture, significantly reducing computational requirements. The results are validated in simulations of ensembles of biological systems, but are generally applicable to ensembles of similar systems.

Baseline: N systems -> N controllers. Proposed: N systems -> K clusters, with K < N.

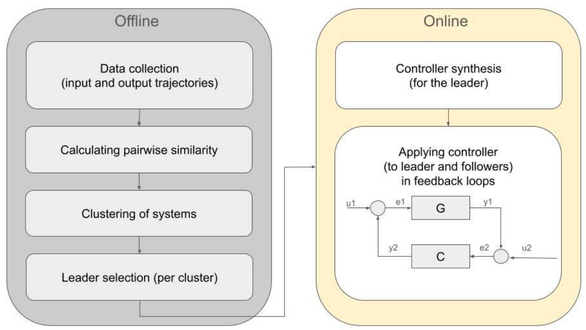
2. Background: Data Driven Control and Cybergenetic Control



Schematic workflow for parallel feedback control of single cells. Cells are isolated and imaged in a mother-machine device; cell-specific measurements are processed by a control computer, which computes light inputs projected onto the cells via a digital micromirror device. The example input is from [1].

In the control loop, input and output data are collected for derivation of control law. No explicit model identification is needed.

3. Proposed pipeline



Offline: collect data, compute pairwise similarity, cluster systems, select leaders. Online: design/apply one controller per leader.

4. Similarity-Aware, Data-Based Clustering

Step A - represent behaviour of a system

Construct Hankel matrices from input-output trajectories, avoiding explicit system identification. The column space of a Hankel matrix with persistently exciting data represents the possible trajectories of the system.
Hence, the behaviour of a system B = col(H).

Step B - find similarity

L-gap formula[1]: gap_L(B1|L, B2|L) = || P1 - P2 ||

Small L-gap means similar finite-length behaviours, making controller reuse more plausible.

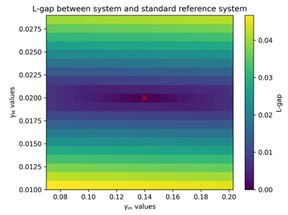


Figure on the left shows the L-gap between reference system (red) and systems with different mRNA and protein dilution/degradation rates, for a simple and linearised biological model. The units are scaled to min^(−1).

Step C - cluster and reuse control

Cluster and choose cluster leaders using pairwise L-gap distances.
Verify followers against a well-posedness threshold.
Design data-driven LQR[2] only for each leader and reuse within the cluster.

This is control-oriented clustering: similarity can be used to evaluate closed loop performance , not just how data look alike.

5. Closed-loop Performance Analysis

Theorem *Let C be the controller derived from system $\mathcal{B}_1|_L$ and assume that $[\mathcal{B}_1|_L, C]$ is well-posed. The closed-loop system $[\mathcal{B}_2|_L, C]$ is well-posed if $\mathrm{gap}_L(\mathcal{B}_1|_L, \mathcal{B}_2|_L) < \varepsilon_L$ where*

$$\varepsilon_L = 1 - \mathrm{gap}_L(\mathcal{B}_1|_L, \mathcal{G}_L(C^\star)) < 1.$$

This checks that reusing the controller on the different system leads to mathematically valid trajectories.

Proposition *The performance cost difference can be bounded by $|J(w_2) - J(w_1)| \le \eta_J \|e\|^2$, where η_J is defined as*

$$\eta_J := \eta_w \|W\| \left(\frac{1}{\sqrt{1-\delta^2}} + \frac{1}{\sqrt{1-(\delta + D_{12})^2}} \right),$$

where η_w is defined as

$$\eta_w := \frac{D_{12}}{\sqrt{1-\delta^2}} \frac{\sqrt{2-(\delta + D_{12})^2}}{\sqrt{1-(\delta + D_{12})^2}},$$

and $D_{12} = \mathrm{gap}_L(\mathcal{B}_1|_L, \mathcal{B}_2|_L)$ and $\delta = \mathrm{gap}_L(\mathcal{B}_1|_L, \mathcal{G}_L(C^\star))$.

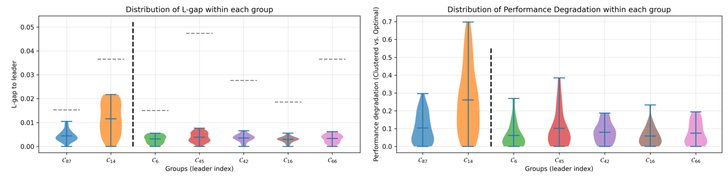
This ensures that the trajectory from [B2, C] remains close to [B1, C] and hence is still stable.

6. Biological test case

The framework was tested on simulated cybergenetic gene-expression systems with parameter heterogeneity.

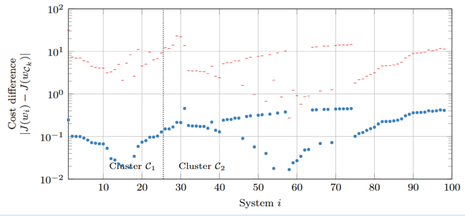
100 systems | **K = 2 or 5** cluster cases | **LQR** data-driven

Clustering Results and Comparison to Optimal Trajectories



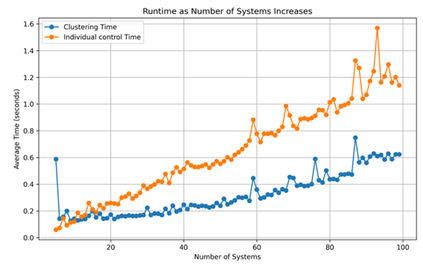
These are the violin plots showing that systems within a cluster are similar and performances are still close to optimal with shared controllers.

Bound Check



This shows that the performance cost differences between leaders and followers remain below the derived bound in the two-cluster case.

Runtime Comparison



This shows that the runtime of clustered control (blue) is better than standard control (orange) on a standard laptop (Intel Core i7 processor, 8 GB RAM)

7. Conclusions

This work demonstrates that dynamical similarity between biological subsystems can be quantified directly from Hankel matrix data using subspace-based L-gap metrics, and that subsystems clustered according to this metric can be effectively controlled through a shared control architecture. The proposed framework significantly reduces computational requirements while maintaining regulation performance, thereby enabling scalable cybergenetic control in high-throughput microfluidic platforms. Although validated here on linearised low-dimensional models, the framework is general and can be extended to higher-dimensional systems and alternative data-based control schemes. Beyond engineering biology, the methodology is applicable to any ensemble of dynamically similar systems where data-driven control and computational efficiency are essential.

References

[1]Jessica James, Idris Kempf, Kirill Sechkar, Jingyu Wang, Gabriel Abrahams, Sebastian Towers, and Harrison Steel. Microscopic photoselection (mips) of single cells in mother machine microfluidic devices. bioRxiv, 2025.

[2]Alberto Padoan, Jeremy Coulson, Henk J. van Waarde, John Lygeros, and Florian Dörfler. Behavioral uncertainty quantification for data-driven control. In IEEE 61st Conference on Decision and Control (CDC), pages 4726–4731, 2022.

[3]Claudio De Persis and Pietro Tesi. Formulas for data-driven control: Stabilization, optimality, and robustness. IEEE Transactions on Automatic Control, 65(3):909–924, 2020.