Journal of Machine Learning Research (2022) Submitted ; Published

Multi-Information Source Bayesian Optimization of Culture Media for Cellular Agriculture

Zachary Cosenza ZACOSENZA@UCDAVIS.EDU
Department of Chemical Engineering
University of California Davis

Raul Astudillo RA598@CORNELL.EDU
Operations Research and Information Engineering
Cornell University

Peter Frazier PF98@CORNELL.EDU
Operations Research and Information Engineering
Cornell University

Keith Baar KBAAR@UCDAVIS.EDU
Departments of Neurobiology, Physiology, and Behavior and Physiology and Membrane Biology
University of California Davis

David E. Block DEBLOCK@UCDAVIS.EDU
Department of Chemical Engineering
University of California Davis

Editor: Leslie Pack Kaelbling

Abstract

Media used in the field of cellular agriculture is difficult to optimize due to the lack of mathematical models of population-level muscle cell growth. When measured in lab, growth assays are convenient but inaccurate, while robust measures of cell number can be time-consuming. In this work, we addressed these difficulties by optimized a cell culture media with 14 components using a multi-information source Bayesian optimization algorithm that locates optimal media conditions based on an iterative refinement of an desirability function. As a model system, we utilized murine C2C12 cells, using AlamarBlue, LIVE stain, and trypan blue exclusion cell counting assays to determine cell number. We were able to design media with 181% more cells than a common commercial variant at parity economic cost, while doing so in 38% fewer experiments than an efficient design-of-experiments method. The optimal medium generalized well to long-term growth, indicating the assay fusion method improved measurement robustness relative to rapid growth assays alone.

Keywords: cellular agriculture, Bayesian optimization, multi-information source, media optimization

1. Introduction

Every bioprocess where cells are used in production requires suitable culture conditions for cell growth. In the cellular agriculture, where cells are grown for consumption to replace carbon-intensive and often unethical animal agriculture, cost-effective media has been identified as the most critical aspect in scale-up and commercialization (O'Neill et al. (2021)). Optimizing these conditions is difficult due to the large number of media components with nonlinear and interacting effects (Brunner et al. (2010)). Additionally, experiments are difficult to conduct at scale due to the expense of laboratory materials and time required to grow cells. This is especially the case when optimizing adherent cell lines used in cultivated meat production because cells must be sub-cultured / passaged, and thus exhibit drastically different dynamics depending how many times the cells have been passaged (Cosenza et al. (2021)).

In this work we address these issues, particularly the challenge of difficult-to-collect data on multi-passage growth (call such data IS_0), by supplementing it with faster but biased approximations of long-term growth (call these IS_i). The advantage of collecting IS_i is that it can be more numerous and highlight regions of the design space of interest to be further considered by IS_0. We then fuse these information sources using a multi-information source (IS) Gaussian process (GP) model. With this statistical model, we can use Bayesian optimization (BO), specifically the multi-point q-expected improvement acquisition function, to select optimal experiments parameterized by desirability function $D(x)$. Whether to sample a point using IS_0 or some IS_i is determined using a combinatorial heuristic that quantifies the $D(x)$-information value of a given set of experiments and allocates IS accordingly. We will show that this BO method is superior to a traditional design-of-experiments (DOE) method and the commercial dulbecco's modified eagle medium (DMEM).

2. Methods

2.1 Experimental

Our model system is the multi-passage growth of C2C12 cells. These are adherent (growing on the culture dish) cells that require some combination of media components listed in Table 1. They are also immortalized, allowing them to be stably passaged multiple times without much genetic drift or cell death. The cells are kept at 37°C and 5% CO2 and passaged every 72 hrs in test media depending on the IS. The quality of a given combination of media components were evaluated as follows: IS_0, or cell count after Passage 2, was the most robust measure of cell health because the C2C12 cells would have to survive and grow for 144 hrs and withstand two passages, which can be damaging to cells. IS_1, or cell count after Passage 1, is "automatically" pared with IS_0 because in order to passage a cell population at a given density one must have the cell count. IS_2 (AlamarBlue) and IS_3 (LIVE) are chemical assays done in small 96 well plates that are correlates of short-term cell health. Together, these assays form the IS used in this study.

2.2 Computational

In standard BO, we model a process using a GP characterized by a prior mean $\mu_0 = c$ and covariance matrix $\Sigma(x, x') = \sigma_f^2 exp(-0.5\Sigma_{k=1}^{p}(x_k - x'_k)^2/\lambda_k^2)$. We used the squared

Abrev.	Component	Conc. Min (mg/mL)	Conc. Max (mg/mL)	Cost c_k
T	Transferrin	0	0.026	6.53E-03
I	Insulin	0	0.035	1.43E-02
SS	Sodium Selenite	0	1.75E-05	6.4E-09
AA	Ascorbic Acid	0	8.75E-03	9.8E-06
Glu	Glucose	0	15.75	0.2
Gluta	Glutamine (GlutaMAX)	0	1.519	2.09E-02
Albu	Albumin (AlbuMAX)	0	1.4	4.94
FBS	FBS (% v/v)	0	17.5	14.00
H	Hydrocortisone	0	1.75E-05	1.1E-05
D	Dexamethasone	0	7.00E-04	7.2E-03
P	Progesterone	0	1.75E-05	4.0E-07
Esd	Estradiol	0	8.75E-06	1.6E-06
Ethan	Ethanolamine	0	6.65E-03	6.1E-06
Glutath	Glutathione	0	3.50E-03	6.0E-04
-	DMEM Supplement (% v/v)	-	***54.3	2.1E-02

Table 1: The bounds of optimization are listed above. The cost shown is a unitless scalarization of the relative economic cost of each component. ***All media have a 54.3% v/v (volume percent) base of DMEM supplement (liquid form, no glucose, glutamine, or FBS).

exponential covariance function covariance kernel to encode the belief that (i) media that are closer in concentration are closer in growth rate, governed by hyper-parameters σ^2 and λ^2, (ii) that the overall biological processes underlying the response surface are smooth with (iii) each component response governed by λ_k, allowing each component k to have different degrees of "wigglyness" for each IS. After observing N data points from a generative process $y(x) = g(x) + \epsilon$ with noise $\epsilon \sim N(0, \sigma_\epsilon^2)$ we can compute the posterior mean $\mu(x)$ and covariance $\sigma(x)$ (equations 2.23 and 2.24 in (Rasmussen and Williams)).

Because the key objective of this work is to optimize an underlying (and data-poor) process IS_0, we fuse different IS by using a GP described in (Poloczek et al. (2017)). This multi-IS GP utilizes auxiliary information sources to model an underlying "true" function. We chose this model over the more typical multi-task GP to encode the prior belief that the generative model includes an underlying "true" function and several biased / variable but correlated auxiliary functions, and to provide the flexibility of allowing different length-scale hyper-parameters λ_k for each IS to be learned from the data. Let us assume a generative model $y = g(x) + \delta(x, m) + \epsilon$ for a given medium combination x at an IS indexed by m. We therefore have one independent GP for the underlying function $g(x)$ and one for each auxiliary IS deviation function $\delta(x, m)$ for the mth auxiliary IS (where $m = 0$ references IS_0).

$$\Sigma(x_m, x'_l) = \Sigma_0(x, x') + 1_{(m \neq 0)} 1_{(m=l)} \Sigma_l(x, x') \tag{1}$$

With this multi-IS GP characterized by the information fusion kernel in we now can make predictions for IS_0 using multiple data sets. Our objective function was a desirability function $D(x)$ (Akteke-Ozturk et al. (2018)) using a media cost function $c(x) = c_{min} + \Sigma_k^p c_k x_k$ and feasibility indicator metric $\phi(x) = 1_{\mu(x) \geq y_L}$.

$$D(x) = \phi(x)\sqrt{\bar{\mu}(x)\bar{c}(x)} \tag{2}$$

The desirability function (i) scales $\bar{\mu}(x) = \frac{\mu(x) - y_L}{y_H - y_L}$ to favor higher growth media where $y_L = 0.5$ and $y_H = 2$, (ii) scales $\bar{c}(x) = \frac{c(x) - c_H}{c_{min} - c_H}$ to favor lower cost media for $c_H = c_{min} + \Sigma_k^p c_k$, and (iii) down-weights media that fails to be $\mu(x) \geq y_L$, or predicted to perform at least equal to or better than a user-defined lower bound y_L. All lab measurements of cell growth (at all IS) are made relative to a control medium (DMEM) so y is a normalized value. We then use $D(x)$ in the multi-point expected improvement acquisition function $\alpha(X) = E[(max\{D(X)\} - D^*(X_N))^+]$ from (Wang et al. (2020)) where $max\{D(X)\}$ is the optimal $D(x)$ and $D^*(X_N)$ is the previous best desirability found from the N observed data points. To acknowledge the fact that data collected in biological experiments are noisy, we refine the acquisition function to be the noisy multi-point expected improvement (Letham et al. (2019)) by additionally sampling (using $R = 2000$ monte-carlo samples, where Σ is the summation operator) previous values $D(X_N)$. Note that the r subscript means rth sample of the objective function using the reparameterization trick.

$$\alpha(X) = 1/R\Sigma_{r=1}^R[(max\{D_r(X)\} - max\{D_r(X_N)\})^+] \tag{3}$$

We now must determine which experiments should be collected with our high-fidelity IS_0 versus lower-fidelity IS. After hyper-parameter optimization using L-BFGS-B to maximize the log-likelihood of the multi-IS GP (with normal priors on λ's and σ_f's, and a gamma prior on σ_ϵ), we solve $X^* = argmax\, \alpha(X)$ using the IS_0 prediction using multi-start L-BFGS-B for $q = 10$ total experiments (the capacity of our lab). Next, we compute $\alpha(X)$ for all $\binom{q}{q_0}$ combinations of X^* for $q_0 = 3$ (IS_0 capacity of our lab). The highest scoring combination of q_0 experiments in X^* is allocated to IS_0 and IS_1, and the $q - q_0 = 7$ remaining experiments are allocated to IS_2 and IS_3 only. The reasoning behind this heuristic is that the most optimal experiments should be allocated the most important IS, while the sub-optimal (but still valuable) experiments should be allocated the remaining auxiliary IS.

3. Results

We compared this algorithm to a DOE method, which only had access to IS_1 AlamarBlue assay, in the optimization of cell culture media. Our BO method found $D(x)$ 132% higher than DOE using 38% fewer experiments (but with 34% higher cost) and a 113% improvement over the control DMEM (with 1.6% higher cost) as seen in Figure 1. Our BO method found that Transferrin, Glutamine, Progesterone, and Estradiol should be at a high relative concentration. Ascorbic Acid, Hydrocortisone, and Dexamethasone should be at a low/zero concentration. As expected, there was a trade-off between number of cells $y(x)$ and medium cost $c(x)$ captured in Figure 1b-1c. More nutrients, especially FBS, improved cell number at the expense of higher cost; this trend then breaks down as more FBS and Albumin have a deleterious effect on growth / desirability. This may be due to the redundancy of

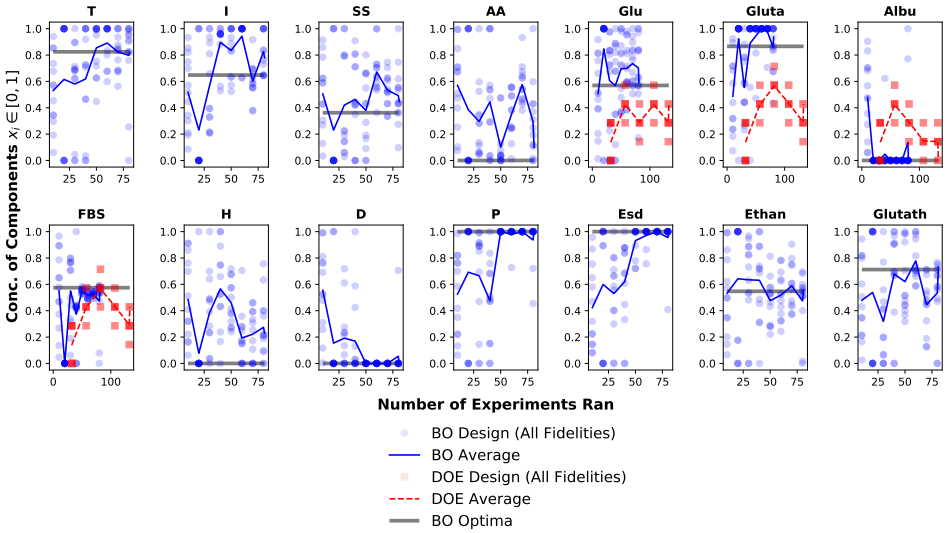

Figure 1: the conditions of each experiment (concentration ranges in Table 1) are shown plotted as a function of the cumulative number of experiments in the BO (circle) and DOE (box) study. The moving average (solid and dashed line for BO and DOE respectively) shows how each method searches for optimal concentrations. The horizontal line represents the final BO optimal concentration.

Albumin (FBS has naturally-occurring Albumin and other proteins) which caused it to be automatically screened out of the design.

It was also useful to examine the correlations between different IS (Figure 2). The model predicts all IS to have very linear correlations (Figure 2c left) likely due to prior-enforced hyper-parameter regularization. In reality, IS_1 and IS_2 fail to capture high performing media due to the breakdown in linearity of the AlamarBlue and LIVE assay at high cell concentrations. This further emphasizes the need for robust measures of multi-passage growth beyond simple (and biased) rapid growth assays. The optimal medium allowed the C2C12 cells to grow well for two additional passages (Passage 4) while the DOE-designed media did not (data not shown), indicating the multi-information source assay improved measurement robustness relative to rapid growth assays alone.

4. Conclusions

Multi-passage growth assays are difficult to measure, and even more difficult to optimize when given many components. We managed this complexity by coupling long-term (> 1 passages) cell number measurements with simpler but less valuable rapid growth chemical assays in murine C2C12 cultures as a model system for cellular agricultural applications, capturing a more holistic model of a theoretical cell growth process. We combined this with an optimization algorithm that efficiently allocates laboratory resources during multi-passage optimization. This resulted in a reduction in experimental effort to find a media

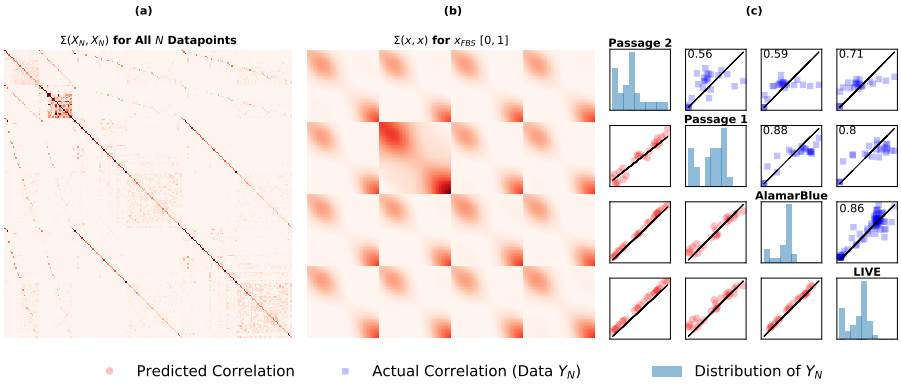

Figure 2: (a) kernel output for data $\{X_N, Y_N\}$ organized by IS_0,IS_1,IS_2,IS_3 left to right (b) same kernel but organized with equally space values of x_{FBS} from 0 to 1. (c) Various IS cell number / correlate distributions (diagonal histograms) are shown. Above the diagonal (squares) are the actual inter-IS correlations for each IS with their respective R2 values, and below the diagonal (circles) are the predicted inter-IS correlations for a random data set. The middle distributions are the output distributions of each IS.

more proliferative than the control at nearly the same cost. As the longer-term passaging study suggests, our Passage 2 objective function and IS were well calibrated to mimicking the complex industrial process of growing large batches of cells over many passages. The media resulting from the BO algorithm supported significantly more C2C12 cell growth with only a small increase in cost. With these results, it should be possible to implement this type of experimental optimization algorithm to other systems of importance to cellular agriculture and cell culture production processes with difficult-to-measure output spaces, including for optimization of serum-free media for cell growth and for differentiation.

Acknowledgments

We would like to acknowledge support for this project from New Harvest Inc.

References

Basak Akteke-Ozturk, Gulser Koksal, and Gerhard Wilhelm Weber. Nonconvex optimization of desirability functions. *Quality Engineering*, 30(2):293–310, 2018. ISSN 15324222. doi: 10.1080/08982112.2017.1315136. URL https://doi.org/10.1080/08982112.2017.1315136.

Daniel Brunner, Helmut Appl, Walter Pfaller, and Gerhard Gstraunthaler. Serum-free Cell Culture : The Serum-free Media Interactive Online Database. (December 2009):53–62, 2010.

Zachary Cosenza, David E Block, and Keith Baar. Optimization of muscle cell culture media using nonlinear design of experiments. *Biotechnology Journal*, 16(11):2100228, nov 2021. ISSN 1860-6768. doi: 10.1002/biot.202100228. URL https://onlinelibrary.wiley.com/doi/10.1002/biot.202100228.

Benjamin Letham, Brian Karrer, Guilherme Ottoni, and Eytan Bakshy. Constrained Bayesian optimization with noisy experiments. *Bayesian Analysis*, 2019. ISSN 19316690. doi: 10.1214/18-BA1110.

Edward N. O'Neill, Zachary A. Cosenza, Keith Baar, and David E. Block. Considerations for the development of cost-effective cell culture media for cultivated meat production. *Comprehensive Reviews in Food Science and Food Safety*, 20(1):686–709, 2021. ISSN 15414337. doi: 10.1111/1541-4337.12678.

Matthias Poloczek, Jialei Wang, and Peter I. Frazier. Multi-information source optimization. In *Advances in Neural Information Processing Systems*, 2017.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*.

Xiaokang Wang, Navneet Rai, Beatriz Merchel Piovesan Pereira, Ameen Eetemadi, and Ilias Tagkopoulos. Accelerated knowledge discovery from omics data by optimal experimental design. *Nature Communications*, 11(1), 2020. ISSN 20411723. doi: 10.1038/s41467-020-18785-y. URL http://dx.doi.org/10.1038/s41467-020-18785-y.