Cytovance Biologics CHO Titer Experiment Analysis

Correlation Matrices and Data Quality Report

Jackson Polk

Summer 2023

DSA 5900-995

4 Credit Hours

Supervisors: Matthew Beattie and Talayeh Razzaghi

Company: Cytovance Biologics

Preface:

I have written most code in R as a “scratchpad”. I will go back and re-write in Python once we have verified that these methods are correct.

Literature Review:

Temperature shift during cell cultivation affects both long and short run production; higher temperatures cultivate quickly but die faster, opposite true for lower temperatures (Xu et. al. 2019). Concludes that temperature is an important factor in cultivation, should be focused on during analysis and modeling.

Data Exploration:

Data Quality Reports

Statistics about each feature/variable can be found here, along with descriptions about missing data.

CHO-KC

Variable	Mean	Median	"Skew"	Variance	# NA	% NA
vessel_name	13.276	13	1.021	45.808	456	0.46
production_day	6.491	6	1.082	18.092	96	0.097
do	0.461	0.5	0.921	0.006	95	0.096
ph_setpoint	7	7	1	0	95	0.096
temp	36.786	37	0.994	0.811	95	0.096
target_cell_seeding	432107	3.00E+05	1.44	64140979605	95	0.096
feed	1.122	0.03	37.385	2.068	212	0.214
glucose_trigger_limit	3.2	3	1.067	0.06	94	0.095
viable	7618971	8285000	0.92	2.43879E+13	144	0.145
cell_viability	94.706	98.1	0.965	131.071	144	0.145
average_cell__diameter	17.718	17.485	1.013	3.531	198	0.2
ph	7.153	7.112	1.006	0.05	129	0.13
titer_by_octet	1569.66	1571.187	0.999	977464.054	860	0.867
glutamine	4.621	5.26	0.878	6.32	683	0.689
glutamate	7.486	6.955	1.076	14.142	404	0.407
glucose	4.778	4.785	0.999	1.531	106	0.107
lactate	1.29	1.25	1.032	0.457	114	0.115
ammonium	4.088	3.87	1.056	3.866	162	0.163
sodium	141.611	134.6	1.052	2103.782	143	0.144
pottasium	4.317	4.28	1.009	3.94	142	0.143
calcium	0.113	0.11	1.023	0	142	0.143
osmolality	534.63	512	1.044	25417.055	689	0.695
bicarbonate	10.262	8.6	1.193	64.813	528	0.532
air_saturation	63.553	68.35	0.93	272.115	816	0.823
co2_saturation	2.525	2.3	1.098	2.91	816	0.823

Fig 1. Numeric Data Quality Report for CHO-KC data. Note: "Skew" is the ratio of Mean to Median, indicating if the data is skewed.

Factor	Levels	# NA	% NA
vessel_type	3	96	0.097
supplement	1	992	1
media	2	95	0.096
feed_type	2	95	0.096
feeding	2	515	0.519
notes	26	456	0.46

Fig 2. Analysis of Factor Variables for CHO-KC data.

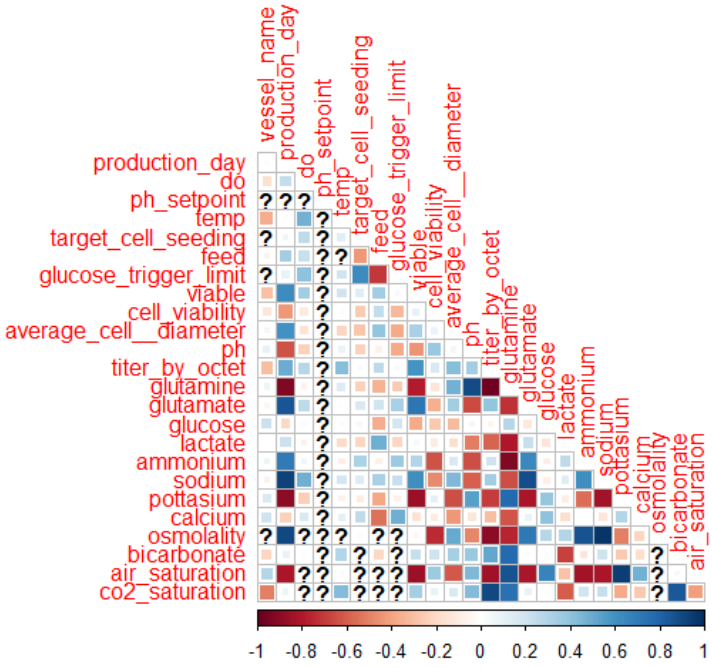

Fig 3. Correlation Matrix for CHO-KC data.

CHO-S

Variable	Mean	Median	"Skew"	Variance	# NA	% NA
production_day	5.612	5	1.122	14.806	92	0.078
do	0.366	0.3	1.219	0.01	531	0.45
ph_setpoint	7.071	7	1.01	0.009	507	0.43
temp	35.477	37	0.959	5.122	95	0.081
target_cell_seeding	1652217	1.00E+06	1.652	1.4255E+12	75	0.064
feed	1.915	3	0.638	1.772	114	0.097
glucose_trigger_limit	4.196	4.5	0.932	0.357	114	0.097
viable	8797862	9130000	0.964	2.16664E+13	114	0.097
cell_viability	92.123	96.4	0.956	138.319	105	0.089
average_cell__diameter	14.966	14.3	1.047	3.808	113	0.096
ph	7.109	7.063	1.006	0.071	388	0.329
titer_by_octet	2088.93	1401.746	1.49	4820087.294	632	0.536
glutamine	2.998	1.54	1.947	8.556	700	0.593
glutamate	5.287	4.735	1.117	5.825	370	0.314
glucose	5.036	4.64	1.085	8.703	295	0.25
lactate	0.953	0.75	1.271	0.754	361	0.306
ammonium	7.116	5.9	1.206	26.184	364	0.308
sodium	105.363	102	1.033	637.184	359	0.304
pottasium	3.652	3.51	1.04	2.007	359	0.304
calcium	0.074	0.07	1.052	0.001	358	0.303
osmolality	351.216	312.5	1.124	11058.669	642	0.544
bicarbonate	9.22	6.4	1.441	46.065	482	0.408
air_saturation	41.064	42.7	0.962	446.566	799	0.677
co2_saturation	3.636	3.5	1.039	1.071	799	0.677

Fig 4. Numeric Data Quality Report for CHO-S data. Note: "Skew" is the ratio of Mean to Median, indicating if the data is skewed.

Factor	Levels	# NA	% NA
vessel_type	3	0	0
vessel_name	27	440	0.373
supplement	9	740	0.627
media	2	0	0
feed_type	3	0	0
feeding	2	216	0.183
notes	33	580	0.492

Fig 5. Analysis of Factor Variables for CHO-S data.

Notice that "vessel_name" appears in this dataset, but not in the CHO-KC data.

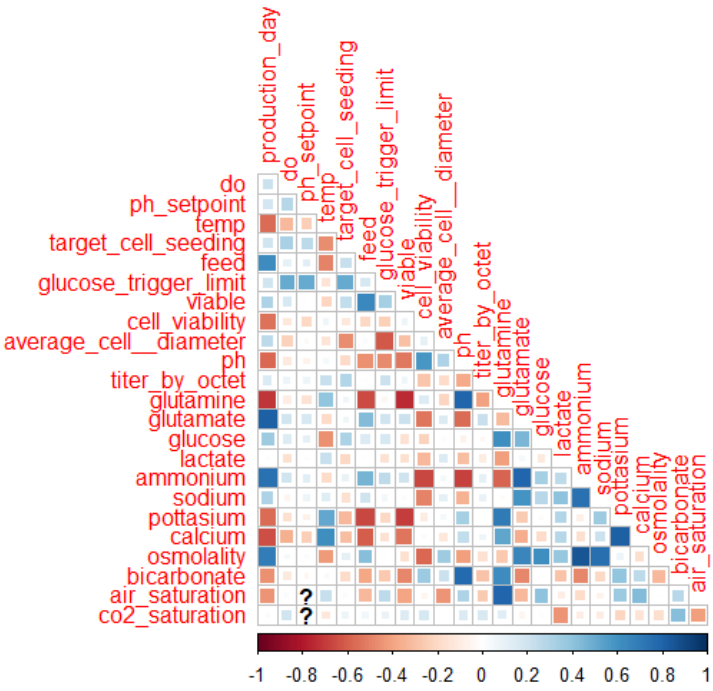

Fig 6. Correlation Matrix for CHO-S data.

Other Data Visualizations

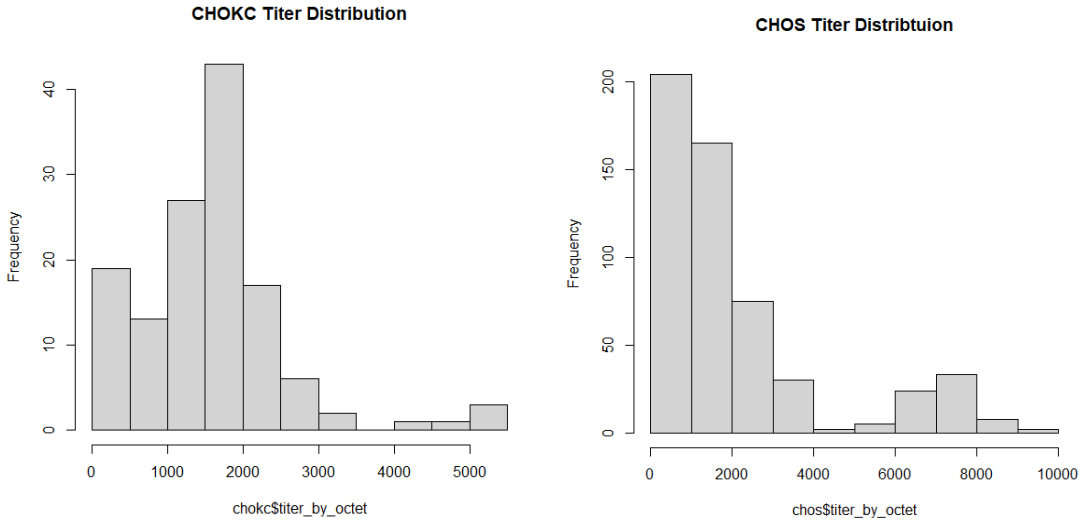

Fig 7. Histogram of Titer from CHO-KC and CHO-S. CHO-KC displayed left, CHO-S right.

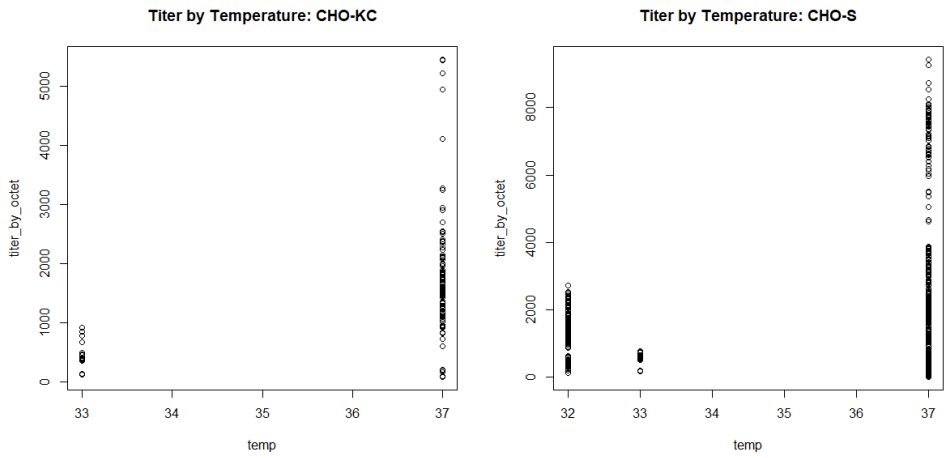

Fig 8. Titer by Temperature. Temperature appears discontinuous.

Although temperature has been suggested to be an important variable in cultivation (Xu et. al., 2019), there is little evidence that Cytovance intentionally manipulated temperature.

Algorithm Proposal:

My initial plan is an online, semi-supervised regression model evaluated using root mean-squared error. The use of a regression model assumes that the exact value of titer is needed, rather than a range. The online model is necessitated by the manufacturing process being continuous, it would be helpful to train the model on new data as it comes in. The semi-supervised learning is due to the sparsely-labelled dataset. Finally, the RMSE method will allow proper weighting of outlier predictions.

<u>Addendum</u>

I have been suggested to use reinforcement learning by Dr. Beattie.

Citations

Xu, J., Tang, P., Yongky, A., Drew, B., Borys, M. C., Liu, S., & Li, Z. J. (2019). Systematic development of temperature shift strategies for Chinese hamster ovary cells based on short duration cultures and kinetic modeling. mAbs, 11(1), 191–204. https://doi.org/10.1080/19420862.2018.1525262