2nd International Electronic Conference on Metabolomics
IECM 20–27 November 2017
SCIforum

Conference Proceedings Paper

A poly-omics machine learning method to predict metabolite production in CHO cells

Guido Zampieri[1], Macauley Coggins[2], Giorgio Valle[1] and Claudio Angione[2,*]

[1] CRIBI Biotechnology Centre, University of Padova, viale G. Colombo 3, 35131 Padova, Italy; guido.zampieri@phd.unipd.it
[2] Department of Computer Science and Information Systems, Teesside University, Borough road, TS1 3BA Middlesbrough, UK; c.angione@tees.ac.uk
* Correspondence: c.angione@tees.ac.uk; Tel.: +4401642342681

Academic Editor: name
Published: date

Abstract: The success of biopharmaceuticals as highly effective clinical drugs has recently led industrial biotechnology towards their large-scale production. The ovary cells of the Chinese hamster (CHO cells) are one of the most common production cell line. However, they are very inefficient in producing desired compounds. This limitation can be tackled by culture bioengineering, but identifying the optimal interventions is usually expensive and time-consuming. In this study, we combined machine learning techniques with metabolic modelling to estimate lactate production in CHO cell cultures. We trained our poly-omics method using gene expression data from varying conditions and associated reaction rates in metabolic pathways, reconstructed *in silico*. The poly-omics reconstruction is performed by generating a set of condition-specific metabolic models, specifically optimised for lactate export estimation. To validate our approach, we compared predicted lactate production with experimentally measured yields in a cross-validation setting. Importantly, we observe that integration of metabolic predictions significantly improves the predictive ability of our machine learning pipeline when compared to the same pipeline based on gene expression alone. Our results suggest that, compared to transcriptomic-only studies, combining metabolic modelling with data-driven methods vastly improves the automatisation of cultures design, by accurately identifying optimal growth conditions for producing target therapeutic compounds.

Keywords: CHO cell; Biopharmaceutical; Metabolic modelling; Machine learning; Flux balance analysis.

1. Introduction

Chinese hamster ovary (CHO) cells are widely regarded as one of the most reliable cell types for industrial-scale mammalian protein production. As compared to bacterial cell lines such as those of Escherichia coli, CHO cultured cells are less

productive, much fragile and grow slowly. In turn, this means that the manufacturing methods that facilitate protein production using CHO cell lines are much more expensive and time-consuming. However, heavy interest is put in optimising CHO cell lines as they are required to produce mammalian recombinant proteins.

Recent advances in this context have focused on unraveling the complex biological machinery controlling desirable characteristics of protein synthesis and secretion [1]. While gene expression profiling has proved helpful in past studies, there have been recent efforts to combine genetic data with knowledge of metabolic pathways through the reconstruction of genome-scale metabolic models (GSMMs). GSMMs attempt to describe cellular metabolism *in silico* through gene annotation and stoichiometry associated with reactions and metabolites, as well as with constraints such as upper or lower bounding of metabolic flux rates. Flux balance analysis (FBA) allows to predict the configuration of metabolic reaction fluxes within GSMMs under general growth conditions [2]. Condition-specific GSMMs can be built using a variety of methods and extended FBA pipelines. The idea is to use omic-data available in each condition, and a set of rules to constrain the flux rates of the general-purpose GSMM [20,21].

Metabolic models have recently been reconstructed for CHO-K1, CHO-S, and CHO-DG44 cell lines, along with a general consensus model [3]. These models were useful in quantifying the protein synthesis capacity of these cell lines and revealed that bioprocessing treatments such as histone deacetylase inhibitors' lead to an inefficiency in increasing product yield. FBA can thus reveal the impact of various media and culture conditions on growth and yield of cultured cells, aiding CHO cells bioengineering [3-6]. Moreover, computational estimation of metabolic fluxes can be an asset when experimental data is not available [7].

However, the precision of GSMMs strongly depends on available pathway and biochemical knowledge. Especially when dealing with the complexity of mammalian cells, more advanced computational techniques may be necessary for an effective application to real problems within the bio-processing industry. In particular, machine learning coupled with computational modelling of CHO cells has the potential to effectively elucidate optimal bioengineering steps towards improved production of therapeutic metabolites and proteins [8].

Here we present a new approach integrating machine learning and metabolic modelling for the computational prediction of protein production in CHO cells. We propose to integrate experimental data on the gene level with data generated in silico via a GSMM of CHO cells metabolism within an integrated data-driven framework (Figure 1). We evaluated this approach by a computational validation, estimating the average prediction error in general settings. Importantly, we observe that metabolic predictions coupled with gene expression data can significantly improve estimations of lactate production based solely on gene expression.

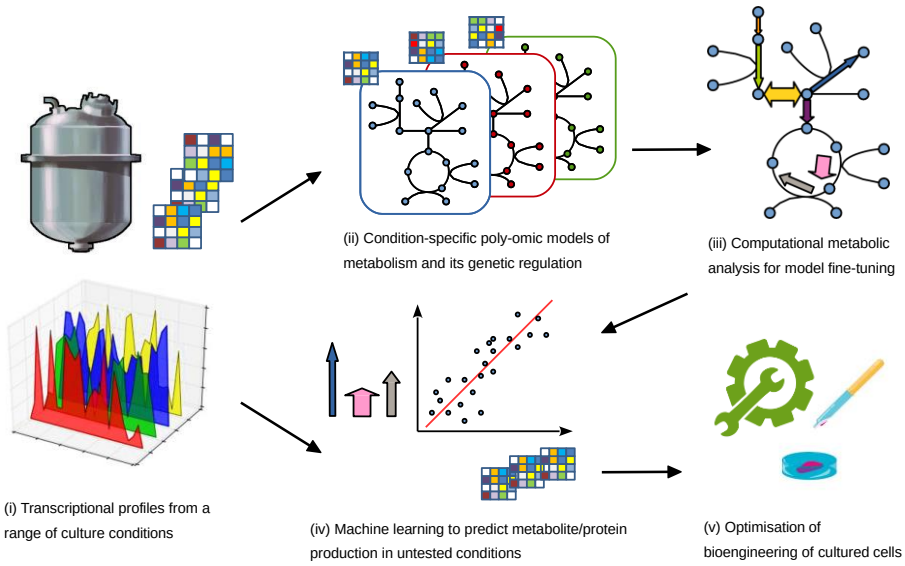

Figure 1. Workflow of the proposed approach for the prediction of metabolite and protein prediction in CHO cells. Steps (i)-(iv) are presented in the Methods section of this work. They serve the final goal of optimising culture bioengineering, depicted in step (v). Integrating transcriptomics data, machine learning methods and metabolic modelling improves the predictive ability of transcriptomic-only methods.

2. Materials and Methods

2.1 Publicly available gene expression data

As a first data source, a large-scale gene expression dataset from two different CHO cell lines was used [9]. This dataset contains 295 microarray profiles with expression values for 3592 genes from 121 CHO cell cultures of varying conditions in terms of including cell density, growth rate, viability, lactate and ammonium accumulation and cell productivity. We extracted the 127 profiles with available quantification of lactate accumulation.

2.2 Genome scale reconstruction of CHO metabolism

We used a recently developed GSMM of CHO cell metabolism, previously used to accurately predict growth phenotypes [3]. This model is the largest reconstruction of CHO metabolism to date, with 1766 genes and 6663 reactions, aggregating community knowledge from various sources. Being a consensus model, it provides general mechanistic relationships that can be refined depending on the particular task or cell line of interest.

2.3 Building condition-specific poly-omics models of CHO cells

To create condition and cell line-specific poly-omics models the genome-scale model of CHO cell metabolism was combined with the gene expression data from CHO cell cultures in varying conditions. In this step, data accessible via the BIGG

repository was employed to match gene identifiers [10]. A model for each condition was created by computing gene set effective expressions Θ for each reaction, following previous investigations [11,12]. The effective expression at reaction level is thereby determined by gene expressions θ(g) and by gene-protein-reaction rules, properly converted to min/max rules depending on the type of gene set. In particular, we define $\Theta(g) = \theta(g)$ for single genes, $\Theta(g_1 \wedge g_2) = \min\{\theta(g1), \theta(g2)\}$ for enzymatic complexes and $\Theta(g_1 \vee g_2) = \max\{\theta(g_1), \theta(g_2)\}$ for isozymes. Lower bounds and upper bounds for each reaction were obtained by applying the following multiplicative coefficient to its native bounds:

$$\phi(\Theta) = [1 + \gamma |\log(\Theta)|]^{sgn(\Theta-1)}, \quad (1)$$

where γ is a parameter controlling the impact of gene expression on reaction bounds.

2.4 Extraction of metabolic features

After a model for each condition was created, flux distributions were computed using FBA by maximising the biomass for producing cell lines included in the CHO model [3]. To perform FBA we employed the COBRA toolbox and a multi-level linear program structure [13,24]. All simulations were carried out in Matlab R2014b with the Gurobi solver.

2.5 Feature processing and selection

Principle Component Analysis (PCA) is a very effective statistical tool that uses an orthogonal transformation to reduce a set of variables to a smaller set of linearly uncorrelated variables, known as the principle components [14]. Here PCA was used to process metabolic flux features in order to extract informative metabolic features.

Moreover, elastic net was applied to select relevant features, both at a gene expression and metabolic level [15]. Given an α in the interval]0, 1] and a non-negative λ, elastic net solves the following optimisation problem:

$$min_{\beta_0,\beta}\left(\frac{1}{2N}\sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2 + \lambda P_\alpha(\beta)\right). \quad (2)$$

In this formula, x represents the gene expression and metabolic flux rates variables, y corresponds to measured metabolite yield and N is the total number of training conditions. $P_\alpha(\beta)$ is a regularisation term depending on a vector of linear coefficients β and on parameter α. Non-null entries of β resulting from this minimisation correspond to relevant features selected by elastic net.

2.6 Training generalised linear models to predict metabolite/protein production

Generalised linear models (GLM) were trained to predict lactate yield starting from poly-omics information [16]. A GLM gives an estimate of metabolite production y_i^{pred} calculated as follows:

$$y_i^{pred} = \beta_0 + x_i^T \beta. \quad (3)$$

GLM accuracy was assessed by nested cross-validation, consisting of two cross-validation loops which together evaluate a selected model based on training data [17]. The nested loop selects the values of α and λ of elastic net on 5 training and test folds. The outer loop is used for model evaluation and is ran over 10 folds. GLM accuracy for each test fold was evaluated by computing the root-mean-square error (RMSE) defined by the following formula:

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i^{pred} - y_i\right)^2}{n}}, \quad (4)$$

where n is the number of test conditions in the fold.

3. Results

3.1. Metabolic model optimisation

We validated our proposed approach on the prediction of lactate production, resorting to experimental data from the study of Clarke et al. [9]. We selected the conditions with both microarray and measured lactate production, obtaining 127 conditions. In order to optimise metabolic flux information, we performed a sensitivity analysis on the gene expression mapping parameter γ in Equation (1). Specifically, we studied the Pearson correlation r between measured lactate accumulation in culture media and simulated lactate export rates for varying values of γ across several orders of magnitude. The maximum correlation coefficient obtained was r = 0.36 (p-value = $2.6 \cdot 10^{-5}$). The relationships between these two quantities can be visualised in Figure 2a. We thus employed condition-specific models with the optimal γ to generate fluxes for the following analysis.

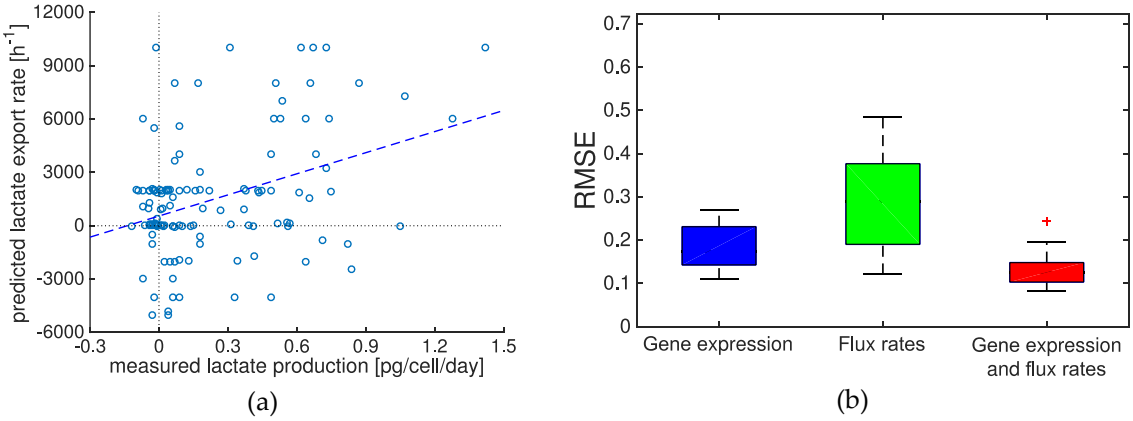

Figure 2. Validation results of the proposed approach on lactate production prediction: **(a)** comparison between simulated lactate export through condition-specific GSMMs and measured lactate production; this step enables GSMMs optimisation for the target metabolite in the following step; **(b)** RMSE distribution plots for lactate production predictions as a function of employed data sources. Two outliers for the green box lie outside of the current scale.

3.2 Predictions of lactate production

To accurately predict lactate production in CHO cells, we employed elastic net and GLMs as described in the Methods section. We estimated the generalised prediction error by means of a 10-fold cross-validation, repeatedly swapping conditions used in training and in tests [17]. We calculated the RMSE of predicted lactate yield across the test conditions in each fold, which quantifies the average difference between predicted and experimentally measured lactate yield. We repeated this procedure under three data sources scenarios, where gene expression, metabolic fluxes and their combination was evaluated separately. The results are shown in Figure 2b and summarised in Table 1. Interestingly, although flux rates alone lead to poor predictions, if combined with gene expression they achieve the minor average and median RMSE across the 10 folds. In the latter case, associated RMSE distribution is significantly different to that obtained from gene expression alone on the basis of a one-tailed Wilcoxon rank sum test at a 5% threshold (p-value = 0.027) [18].

	Gene expression	**Flux rates**	**Gene expression and flux rates**
Mean RMSE	0.19	1.08	0.14
Median RMSE	0.17	0.26	0.13
RMSE standard deviation	0.06	2.41	0.05

Table 1. Comparison of 10-fold cross-validation RMSE statistics for the prediction of lactate production from different data sources. Combining gene expression and metabolic flux data leads to best values for all statistical measures. These results correspond to those shown in Figure 2b.

4. Discussion

The growing demand for natural products in global healthcare requires advanced automation of CHO cell culture design for biotechnological industry to reach commercial-scale production levels. Notably, recent advances in metabolic modelling and in data-driven prediction algorithms have not been yet exploited in combination for this purpose. In this study, we started to explore this research line: the overall goal of the work was to develop a poly-omics approach capable of predicting metabolite/protein production in CHO cells. The approach comprises a GLM trained on gene expression data originating from cultures in varying conditions and on metabolic flux rates obtained *in silico* from FBA on a GSMM of CHO metabolism. The accuracy of our approach was evaluated in comparison to GLMs employing only a single type of data. This allowed us to show that

combining gene expression and metabolic fluxes improves accuracy compared to just using gene expression or metabolic fluxes separately.

Generation of condition-specific metabolic information can in principle be achieved through various types of computational analysis. In this study, we used FBA as this is the most widely used technique to capture flux configurations in a growth steady state [2]. In principle, different techniques could potentially extract even more useful information, further improving final data-driven predictions. For instance, in a preliminary evaluation we tested also a modified version of parsimonious enzyme usage FBA minimising the norm-2 of reaction fluxes [22,23]. However, we observed that normal FBA achieved best results (data not shown).

The main limitation of this work is represented by a scarce availability of large-scale public data on CHO cells and by the prototypical state of present GSMMs. Proposed strategies for model refining are expected to lead to further prediction improvements [19]. With more comprehensive datasets, both in terms of number of samples and in terms of metabolic gene coverage, we expect our pipeline to vastly improve its predictive ability. Moreover, although our validation focussed on lactate production, the proposed methodological framework can be straightforwardly implemented around any target metabolite or protein.

Despite the above-mentioned limitations, our results show that metabolism-based machine learning methods can significantly improve the predictive power of common transcriptomic-only methods. This is due to the introduction of metabolic features coupled with transcriptomic features. The present study therefore represents a preliminary assessment that we plan to extend in future investigations.

Acknowledgments: This work was partially supported by funding from BBSRC/EPSRC BioProNET. We thank Jonathan Welsh from CPI-NBMC for helpful discussions about CHO cell products.

Author Contributions: C.A. and G.Z. conceived and designed the experiments; G.Z. and M.C. performed the experiments; G.Z. analysed the data; C.A. G.Z. and G.V. contributed analysis tools; G.Z., M.C. and C.A. wrote the paper. All authors read and approved the final version of the paper.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Richelle A. and Lewis N.E. Improvements in protein production in mammalian cells from targeted metabolic engineering. *Curr. Opin. Syst. Bio.* **2017**, 6, 1-6, doi:10.1016/j.coisb.2017.05.019.

2. Orth J.D., Thiele I. and Palsson B.O. What is flux balance analysis? *Nat. Biotech.* **2010**, 28, 245-248, doi:10.1038/nbt.1614.

3. Hefzi H., Ang K.S., Hanscho M., et al. A Consensus Genome-scale Reconstruction of Chinese Hamster Ovary Cell Metabolism. *Cell Syst.* **2016**, 3(5), 434-443.e8, doi:10.1016/j.cels.2016.10.020.

4. Martínez V.S., Dietmair S., Quek L.E., Hodson M.P., Gray P. and Nielsen L.K. Flux balance analysis of CHO cells before and after a metabolic switch from lactate production to consumption. *Biotechnol. Bioeng.* **2013**, 10(2), 660-6, doi:10.1002/bit.24728.

5. Rejc Ž., Magdevska L., Tršelič T., Osolin T., Vodopivec R., Mraz J., Pavliha E., Zimic N., Cvitanović T., Rozman D., Moškon M. and Mraz M. Computational modelling of genome-scale metabolic networks and its application to CHO cell cultures. *Comp. Biol. Med.* **2017**, *88*, 150-160, doi:10.1016/j.compbiomed.2017.07.005.

6. Pan X., Dalm C., Wijffels R.H. and Martens D.E. Metabolic characterization of a CHO cell size increase phase in fed-batch cultures. *Appl. Microbiol. Biotechnol.* **2017**, 101(22), 8101-8113, doi:10.1007/s00253-017-8531-y.

7. Sengupta N., Rose S.T. and Morgan J.A. Metabolic flux analysis of CHO cell metabolism in the late non-growth phase. *Biotechnol. Bioeng.* **2011**, 108(1), 82-92, doi:10.1002/bit.22890.

8. Galleguillos S.N., Ruckerbauer D., Gerstl M.P., Borth N., Hanscho M. and Zanghellini J. What can mathematical modelling say about CHO metabolism and protein glycosylation? *Comput. Struct. Biotechnol. J.* **2017**, 15, 212-221, doi:10.1016/j.csbj.2017.01.005.

9. Clarke C., Doolan P., Barron N., Meleady P., O'Sullivan F., Gammell P., Melville M., Leonard M. and Clynes M. Large scale microarray profiling and coexpression network analysis of CHO cells identifies transcriptional modules associated with growth and productivity. *J. Biotechnol.* **2011**, 155(3), 350-359, doi:10.1016/j.jbiotec.2011.07.011.

10. King Z.A., Lu J.S., Dräger A., Miller P.C., Federowicz S., Lerman J.A., Ebrahim A., Palsson B.O. and Lewis N.E. BiGG Models: A platform for integrating, standardizing, and sharing genome-scale models. *Nucleic Acid Res.* **2016**, 44(D1), D515-D522, doi:10.1093/nar/gkv1049.

11. Angione C. and Lió P. Predictive analytics of environmental adaptability in multi-omics network models. *Sci. Rep.* **2015**, 5, 15147, doi:10.1038/srep15147.

12. Angione C. Integrating splice-isoform expression into genome-scale models characterizes breast cancer metabolism. *Bioinformatics* **2017**, btx562, doi:10.1093/bioinformatics/btx562.

13. Schellenberger J., Que R., Fleming R.M.T., Thiele I., Orth, J.D., Feist, A.M., Zielinski D.C., Bordbar, A., Lewis, N.E., Rahmanian S., Kang J., Hyduke D.R. and Palsson B.O. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0. *Nat. Prot.* **2011**, 6(9), 1290-1307, doi:10.1038/nprot.2011.308.

14. Joliffe I.T. *Principal component analysis*, Series: Springer Series in Statistics, 2nd ed.; Springer, New York, United states, 2002.

15. Zou H. and Hastie T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **2005**, 67, 301-320, doi:10.1111/j.1467-9868.2005.00503.x.

16. McCullagh P. and Nelder J.A. *Generalized linear models*, 2nd ed.; Chapman and Hall, London, United Kingdom, 1989.

17. Devijver P.A. and Kittler J. *Pattern Recognition: A Statistical Approach;* Prentice-Hall, London, United Kingdom, 1982.

18. Hollander M. and Wolfe D.A. *Nonparametric Statistical Methods*; Hohn Wiley & Sons, Inc., Hoboken, United States, 1999.

19. Chowdhury R., Chowdury A. and Maranas C.D. Using Gene Essentiality and Synthetic Lethality Information to Correct Yeast and CHO Cell Genome-Scale Models. *Metabolites* **2015**, 29;5(4), 536-70, doi:10.3390/metabo5040536.

20. Vijayakumar S., Conway M., Lió P. and Angione, C. Seeing the wood for the trees: a forest of methods for optimization and omic-network integration in metabolic modelling. *Briefings in Bioinformatics* **2017**, bbx053, doi: 10.1093/bib/bbx053

21. Opdam S., Richelle A., Kellman B., Li S., Zielinski D.C., Lewis N.E.. A Systematic Evaluation of Methods for Tailoring Genome-Scale Metabolic Models. *Cell Systems* **2017**, 22;4(3), 318-29, doi: 10.1016/j.cels.2017.01.010

22. Lewis N.E., Hixson K.K., Conrad T.M., Lerman J.A., Charusanti P., Polpitiya A.D., Adkins J.N., Schramm G., Purvine S.O., Lopez-Ferrer D., Weitz K.K.. Omic data from evolved E. coli are consistent with computed optimal growth from genome-scale models. *Molecular systems biology* **2010**, 6(1):390, doi: 10.1038/msb.2010.47

23. Kim M.K., Lane A., Kelley J.J., Lun D.S. E-Flux2 and SPOT: validated methods for inferring intracellular metabolic flux distributions from transcriptomic data. *PloS one* **2016**, 11(6):e0157101, doi: 10.1371/journal.pone.0157101

24. Angione C., Conway M., Lió P. Multiplex methods provide effective integration of multi-omic data in genome-scale models. *BMC bioinformatics* **2016**, 17(4):83, doi: 10.1186/s12859-016-0912-1