ScFEA: A graph neural network model to estimate cell-wise metabolic flux using single cell RNA-seq data

Norah Alghamdi[1+], Wennan Chang[1,2+], Pengtao Dang[1,2], Xiaoyu Lu[1], Changlin Wan[1,2], Zhi Huang[1,2], Jiashi Wang[1], Melissa Fishel[3], Sha Cao[1, 4]*, Chi Zhang[1,2]*

[1]Department of Medical and Molecular Genetics and Center for Computational Biology and Bioinformatics, [3]Department of Pediatrics, [4]Department of Biostatistics, Indiana University School of Medicine, Indianapolis, IN, 46202, USA.

[2]Department of Electrical and Computer Engineering, Purdue University, Indianapolis, IN, 46202, USA

*To whom correspondence should be addressed. +1 317-278-9625; Email: czhang87@iu.edu. Correspondence is also addressed to Sha Cao, Email: shacao@iu.edu.

[+]These authors have equal contribution to this work.

SUPPLEMENTARY METHODS

Collection and reorganization of human metabolic map

We reorganized the human metabolic network into different reaction types including metabolism, transporter, and biosynthesis. The reorganized network includes 21 super module classes of 175 modules. For the metabolism part, all reactions were collected from Kyoto Encyclopedia of Genes and Genomes database (KEGG) (61). The first super module includes 121 Glucose and TCA cycle reactions. The glycolysis pathway has major out-branches including polysaccharides synthesis, pentose phosphate, serine metabolism, lactate production and acetyl-coA downstream metabolism, hence were split into seven modules. Most of the TCA cycle intermediate substrates are with branches, so the TCA cycle was split into six modules. This super module is regarded as the central metabolism pathway. The main role of this super module are for energy (ATP) production and fueling other metabolic and biosynthesis pathways with acetyl-coA. The second super module is serine metabolism, which contains 220 reactions. This pathway plays a crucial role in controlling the balance and demand of amino acid types [1]. The Pentose Phosphate pathway (PPP) forms the third super module, contains 44 reactions involved in the biosynthesis of PRPP, a precursor for nucleic acids biosynthesis [2]. The fourth super module is biosynthesis and metabolism of fatty acids, which connects the main metabolic map only via the acetyl-coA. The fatty acids biosynthesis and metabolism pathways have a series of parallel reactions chains for different types of fatty acids. This super module contains two modules of fatty acid synthesis and metabolism, totaling 148 reactions [3]. We collected all amino acid metabolic pathways from KEGG database and rebuild super modules based on the network topology. In total, we generated six super modules of amino acids metabolism, namely Aspartate, Beta-Alanine, Glutamate, and Leucine/Valine/Isoleucine metabolism pathways and Urea Cycle. The aspartate metabolism pathway has 16 enzymes catalyzing 37 reactions, B-alanine metabolism pathway includes 21 enzymes carrying 130 reactions, glutamate metabolism pathway is with 10 enzymes and 21 reactions, and 16 enzymes for urea cycle, respectively. Each of the three essential metabolite leucine, isoleucine, and valine, has a separate pathway. Two additional metabolic super modules are Propionyl-CoA metabolism for exchange of multiple coenzyme A types and spermidine metabolism related to the glutathione and S-adenosyl-L-methionine (SAM) metabolisms.

Transporters enable the movement of molecules between two side of cell membranes. We collect human transporter genes and annotations from Transporter Classification Database, by using the symbol and description in this database [4, 5]. We collected 116 transporter genes of 35 metabolites presented in the metabolic and biosynthesis modules.

An essential part of metabolic map is biosynthesis pathways. KEGG database and literature [6-11] are the main information sources used for building biosynthesis modules. We collected 69 biosynthesis modules forming 9 super modules, namely biosynthesis of hyaluronic acid, glycogen, glycosaminoglycan, N-linked glycan, O-linked glycan, Sialic acid, Glycan, Purine and Pyrimidine. Overall, the biosynthesis modules include 142 enzymes catalyzing 280 reactions.

Scalability and Identifiability.

Scalability analysis. The major computational consumption of scFEA comes from the training of multiple neural networks. The activation function has $O(1)$ time complexity. The time complexity for one complete update is $O(e * N * (i * h + h * m))$ for neural networks with three hidden layers, where i is input layer node number, h is hidden layer node number, m is output layer node number, N is the total number of cells, e is the number of iterations. In this study, $i = \sum_{m=1}^{M} i_m = 726$, $h = M \times 8$, $m = 175$, $e = 100$, N is cell number for each dataset.

Identifiability. The number of parameters in the complete model is $\sum_{m=1}^{M} i_m + 8^x M$, where i_m is the number genes in module m ($M = 175$ and $\sum_{m=1}^{M} i_m = 726$ for the complete map) and x is the number of layers of f_{nn}^{m} ($x = 2\ or\ 3$ as the default setting). The number of constraints is the total number of metabolites, K ($K = 84$ for the complete map). Hence the number of total constraints divided by the number of parameters is $\frac{KN}{\sum_{m=1}^{M} i_m + 8^x M}$, which is $0.0238 * N$ and $0.007 * N$ when $x = 2$ or 3. For a scRNA-seq data with $N \sim 10^2$ for the data generated by constructing a library for each individual cell or $N \sim 10^3$ for drop-seq data, selecting $x = 2\ or\ 3$ the $\frac{\#constraints}{\#parameters}$ is much larger than 1, hence guarantees the identifiability and mathematical correctness of the formulation.

SUPPLEMENTARY FIGURES AND TABLES

Supplementary Tables

Supplementary Table S1. Information of reorganized human metabolic map.

Supplementary Table S2. Metabolomics data and clusters of metabolic modules derived in the Pa03c cell line data.

Supplementary Table S3. Differentially expressed genes (DEG) and Pathway Enrichment (PE) results of the Pa03c cell line data.

Supplementary Table S4. Predicted cell type specific fluxome and metabolic imbalance in the melanoma and head and cancer data.

Supplementary Figures

Supplementary Figure S1. The impact of each gene to the metabolic module 1-14 (glycolysis and TCA cycle modules) in the Pa03c cell line data. The x-axis represents genes and y-axis represents impacts. The larger absolute value on the y-axis indicates a stronger impact of the gene to the metabolic module.

SUPPLEMENTARY REFERENCES

1. Mattaini, K.R., M.R. Sullivan, and M.G. Vander Heiden, *The importance of serine metabolism in cancer.* The Journal of cell biology, 2016. **214**(3): p. 249-257.
2. Jin, L. and Y. Zhou, *Crucial role of the pentose phosphate pathway in malignant tumors (Review).* Oncol Lett, 2019. **17**(5): p. 4213-4221.
3. Mikalayeva, V., et al., *Fatty Acid Synthesis and Degradation Interplay to Regulate the Oxidative Stress in Cancer Cells.* International journal of molecular sciences, 2019. **20**(6): p. 1348.
4. Bhutia, Y.D., et al., *SLC transporters as a novel class of tumour suppressors: identity, function and molecular mechanisms.* The Biochemical journal, 2016. **473**(9): p. 1113-1124.
5. Lin, L., et al., *SLC transporters as therapeutic targets: emerging opportunities.* Nature reviews. Drug discovery, 2015. **14**(8): p. 543-560.
6. DeAngelis, P.L., J. Liu, and R.J. Linhardt, *Chemoenzymatic synthesis of glycosaminoglycans: Re-creating, re-modeling and re-designing nature's longest or most complex carbohydrate chains.* Glycobiology, 2013. **23**(7): p. 764-777.
7. Gao, C. and K.J. Edgar, *Efficient Synthesis of Glycosaminoglycan Analogs.* Biomacromolecules, 2019. **20**(2): p. 608-617.
8. Krasnova, L. and C.-H. Wong, *Understanding the Chemistry and Biology of Glycosylation with Glycan Synthesis.* 2016. **85**(1): p. 599-630.
9. Lv, X., et al., *Synthesis of Sialic Acids, Their Derivatives, and Analogs by Using a Whole-Cell Catalyst.* Chemistry (Weinheim an der Bergstrasse, Germany), 2017. **23**(60): p. 15143-15149.
10. Moffatt, B.A. and H. Ashihara, *Purine and pyrimidine nucleotide synthesis and metabolism.* The arabidopsis book, 2002. **1**: p. e0018-e0018.
11. Zulueta, M.M., et al., *Synthesis of glycosaminoglycans*. 2016. p. 235-261.