

MOTL: Multi-omics matrix factorization with transfer learning

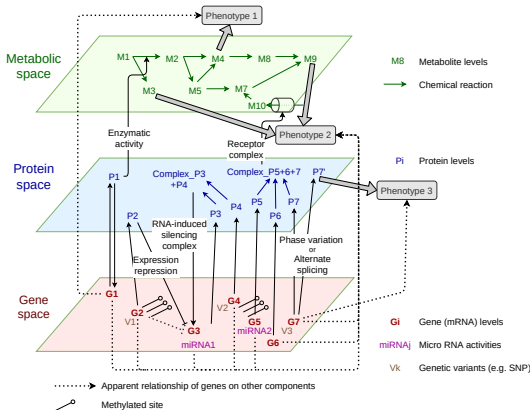
D. Hirst, M. T erezol, L. Cantini, P. Villoutreix, M. Vignes[†] & A. Baudot

School of Mathematical and Computational Sciences
Massey University, NZ

Australasian Applied Statistics Conference, Rottnest Island, WA,
Australia, September 2024

[†] m.vignes@massey.ac.nz

Multi-omics measurement → Joint factorisation and transfer learning



MOTL: Multi-omics matrix factorization with transfer learning
 David Hini¹, Morgane Terzaoui¹, Laura Cantin², Paul Villouhain¹, Matthieu Vignès¹ and Anais Baudin^{1*}

bioRxiv preprint doi: <https://doi.org/10.1101/2023.07.12.552102>; this version posted July 12, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Abstract

Multi-omics → complementary groups of view in M matrices
 Matrix factorisation → lower dimensional representation of data with latent factors associated with underlying biological signals
 Joint multi-factorisation is effective but challenging (dimension, heterogeneity)
 Small sample size issue → Transfer learning
 Usefulness demonstrated on single omics, not on multi-omics

Introduction

a Prior factorization with MOFA

b Transfer learning with MOTL

Key MOTL contributions

Multi-omics target matrices $F = (F^1, \dots, F^M)$ with $F^m \in \mathbb{R}^{n \times p_m}$ and feature weight matrices $W^m \in \mathbb{R}^{n \times k_m}$ with $F^m \approx Z W^m$ and $Z \in \mathbb{R}^{n \times k}$

Also jointly factorise F^m into an sample score matrix, $Z \in \mathbb{R}^{n \times k}$, and feature weight matrices specific matrices, $W^m \in \mathbb{R}^{n \times k_m}$

Hypothesis: If a large L ($n > N$) sampled comprise heterogeneous biological conditions, factorising it with MOFA (Epigat et al., 2016) yields relevant information to factorise F (details in Hini et al., 2024). Algorithm: convergence maximization of \mathcal{L} (maximize), K selected based on fraction of explained variance

We generated $(\times 30)$ multi $M = 3$ omics datasets F (count, continuous and binary), split into target F and learning L sets $F^m \rightarrow \text{Mat}(\text{det}(Z, W^m))$ with $K = 20$ or 30 groundtruth factors. $Z \rightarrow$ based on sample group membership with sampled means (3 values) and same sd

F has 2 groups of 5 samples ($N_1 = 10$)
 L has 20 or 40 groups of random sizes ($2N = 400$ or 1000)
 We compare direct F MOFA to MOTL using F scores

Results

More heterogeneous and aggressive cancer type with sub-groups CL, PH and MS identified in 8 patients (4 healthy)

MS: 2 miRNA expression, DNA methylation. Same L as TCGA (no glioblastoma)

miRNA: 1/8 active factors; MOTL: 10/25 active factors

Heterogeneity of sample clustering based on latent factors →

Conclusion

Epigat et al. (2016) Multi-Omics Factor Analysis framework for integrative multi-omics data sets. Molecular Systems Biology, 12(1) e6555

Hini et al. (2024) MOTL: multi-omics matrix factorization with transfer learning. bioRxiv: <https://doi.org/10.1101/2023.07.12.552102>

Thank you for your attention

See you at the poster session!

Preprint on bioRxiv <https://doi.org/10.1101/2024.03.22.586210> and
Code available on GitHub: <https://github.com/david-hirst/MOTL>