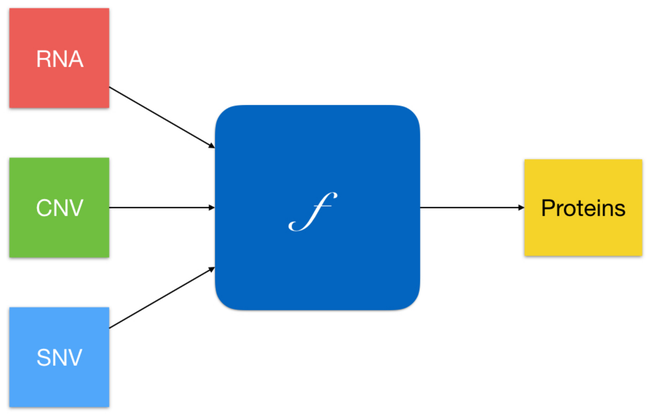
CoDON: a learning framework for linking genomics and transcriptomics data to protein expression

Matteo Manica[1,2], Roland Mathis[1], María Rodríguez Martínez[1]
IBM Research – Zürich, Switzerland[1], ETH Zürich, Switzerland[2]

Outline

Motivation

- **Explain** the **protein expression profiles** through the integration of genomic and transcriptomic data
- **Link different data types** in the context of **complex diseases**
- Wide range of applications (e.g.: multi-omics perturbation analysis)

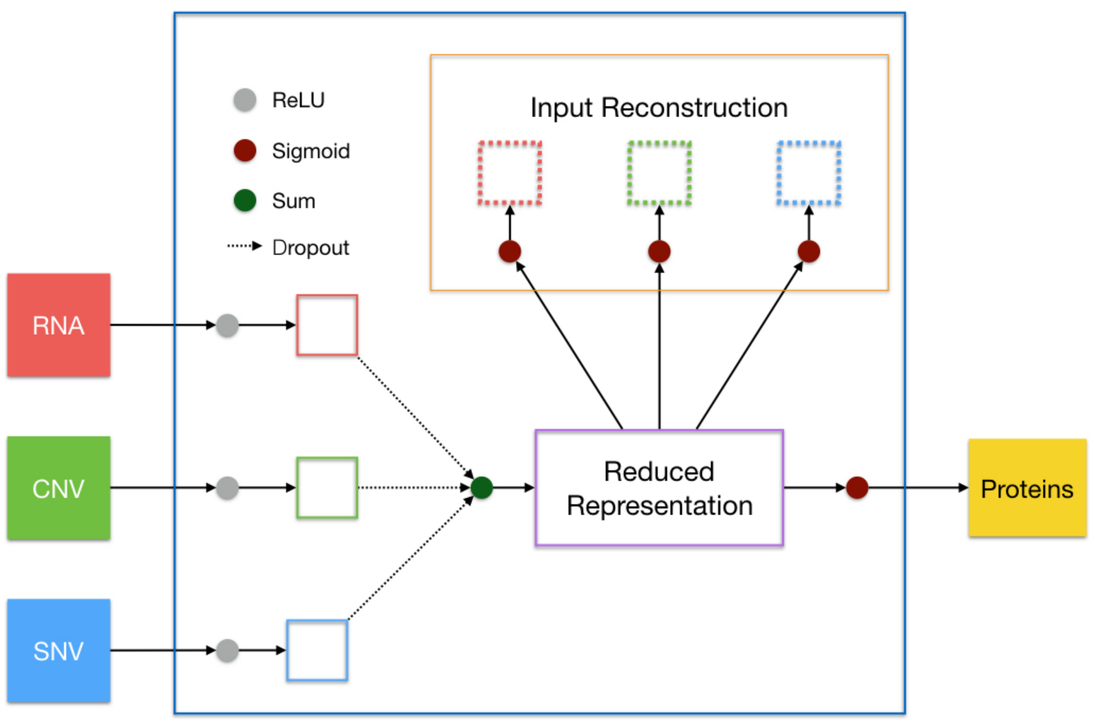


Aims

- **Reconstruct proteins profiles** from different data layers using a **common learning framework**
- **Understand** how different omic datasets can be combined in a **common reduced representation**
- **Decipher** highly **non-linear molecular interactions**

Model

Learning Architecture



CoDON is based on a **neural network architecture**.
A **joint training procedure** is performed, where a lower dimensional representation is built and used to **infer proteomic profiles**, while optimizing the reconstruction on the genomic and transcriptomic level.

Main features:

- Shallow architecture allows to **use weights to analyze complex molecular interactions** within and between different data levels
- **Reconstructing the input datasets** improves the common reduced representation and the reconstruction of the protein profiles[1]
- **Dropout regularization**[1] in merging of the input activation units is used to avoid overfitting

Learning Algorithm

$$\min_{w} \sum_{t \in \mathcal{T}} \|X_t - \mathcal{F}_t(w, X_{RNA}, X_{CNV}, X_{SNV})\|_{L^2}$$

$$\mathcal{T} = \{RNA,\ CNV,\ SNV,\ proteins\}$$

$\mathcal{F}_t(w, X_{RNA}, X_{CNV}, X_{SNV})$ is the non-linear transformation for the reconstruction of dataset t using the learned weights and the input data.

Batch optimization performed using Adam[2] with a fixed learning rate.

Preliminary Results

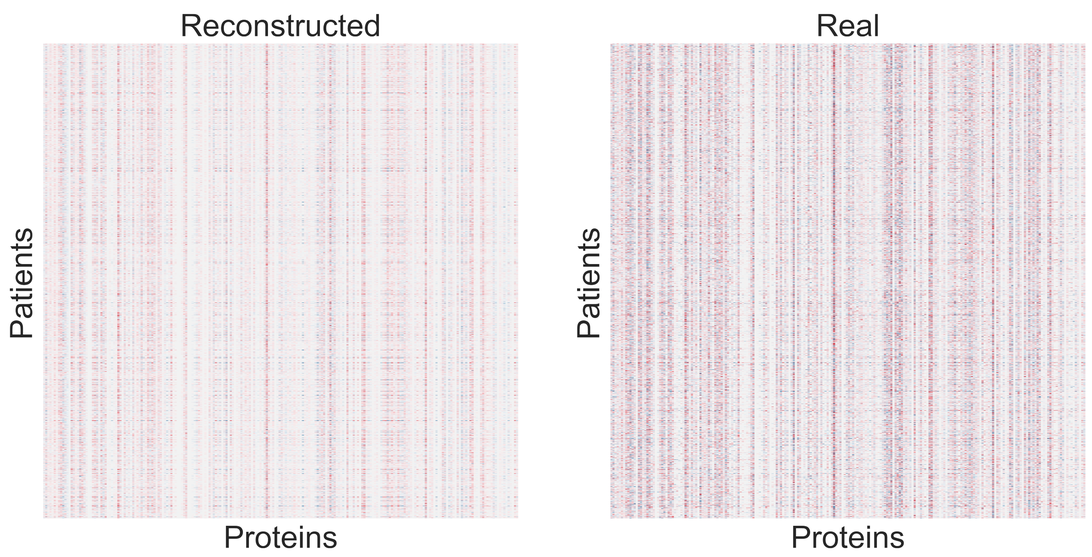
Data

The results were obtained using public data from TCGA[3] for different cancer types:

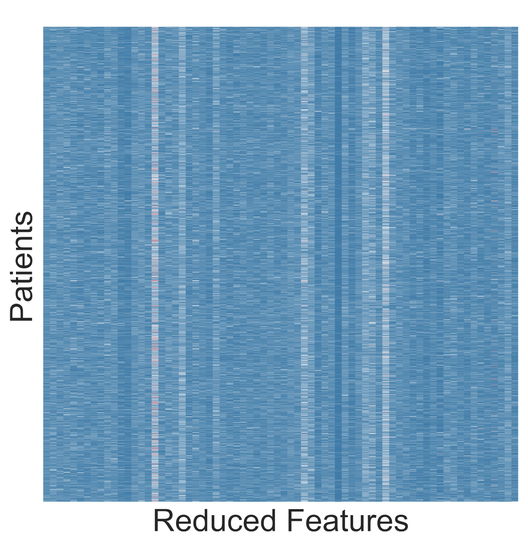
- **RNA-Seq** Version 2 data processed with RSEM[4]
- Gene-wise **CNV** data processed with GISTIC2.0[5]
- Mutation data processed to obtain gene-wise count of **SNV**
- **RPPA** proteins expression data annotated with genes

Patients and genes sets given by the intersection of available datasets were considered (4609 patients and 224 genes).

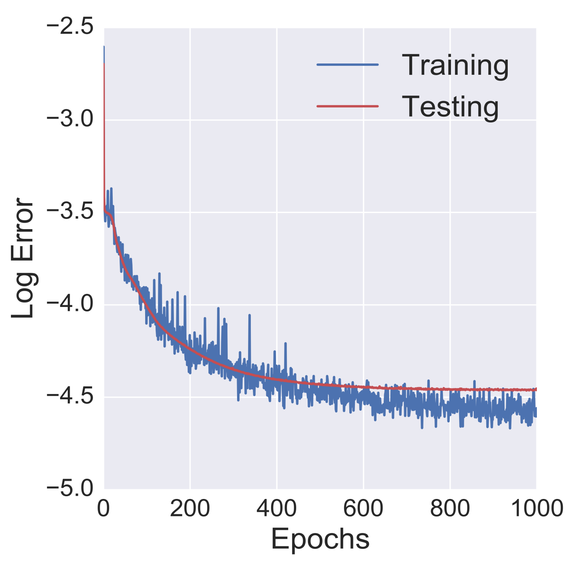
Proteins Profiles Reconstruction



Reduced Representation



Learning Diagnostic



References

[1] Goodfellow, I., Benjo, Y. & Courville, A. Deep Learning. (2016).
[2] Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. CoRR abs/1412.6980, (2014).
[3] TCGA: The Cancer Genome Atlas Research Network, http://cancergenome.nih.gov/.
[4] Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 1–16 (2011).
[5] Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biology 12, 1–14 (2011).

Acknowledgments

CoDON is part of the PrECISE project, a pilot project that combines hypothesis-driven strategies with data-driven analysis in a novel mathematical and computational methodology for the integration of genomic, transcriptomic, proteomic, and clinical data with the goal of risk-stratifying patients and suggesting personalized therapeutic interventions. It has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 668858. We thank The Cancer Genome Atlas (TCGA) Network for granting access to the a vast collection of multi-omics cancer datasets. We also strongly thank IBM Research – Zürich and ETH Zürich for their collaboration in the development of this project.