DreamAI: Algorithm for the Imputation of proteomics data

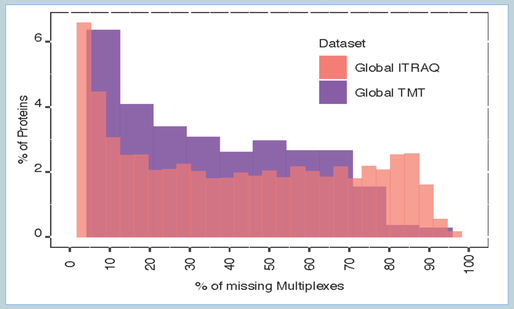
Weiping Ma[1*], Sunkyu Kim[2*]; Shrabanti Chowdhury[1], Zhi Li[3], Mi Yang[4], Seungyeul Yoo[1], Francesca Petralia[1], Jeremy Jacobsen[5], Jingyi Jessica Li[6], Xinzhou Ge[6], Kexin Li[7], Thomas Yu[8], Nathan Edwards[9], Samuel Payne[10], Paul C. Boutros[11], Henry Rodriguez[12], Gustavo Stolovitzky[13], Jun Zhu[1], Jaewoo Kang[2], David Fenyo[3], Julio Saez-Rodriguez[5], Pei Wang[1#]

[1]Icahn School of Medicine at Mount Sinai (USA), [2]Department of Computer Science and Engineering, Korea University (South Korea), [3]New York University (USA), [4]Faculty of Biosciences, Heidelberg University (Germany), [5]University of Colorado (USA), [6]Department of Statistics, University of California Los Angeles (USA), [7]Department of Mathematics, Tsinghua University (China), [8]Sage Bionetworks (USA), [9]Georgetown University (USA), [10]Pacific Northwest National Laboratory (USA), [11]University of California, Los Angeles (USA), [12]National Cancer Institute (USA), [13]IBM Research & Mount Sinai (USA),

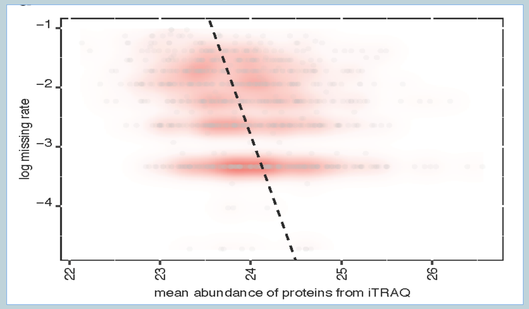
Introduction

Deep proteomics profiling using labelled LC-MS/MS experiments has been proven to be powerful to study complex diseases. However, due to the dynamic nature of the discovery mass spectrometry, the generated data contain a substantial fraction of missing values. This poses great challenges for data analyses, as many tools, especially those for high dimensional data, cannot deal with missing values directly. To address this problem, the NCI-CPTAC Proteogenomics DREAM Challenge was carried out to develop effective imputation algorithms for labelled LC-MS/MS proteomics data through crowd learning. The final resulting algorithm, **DreamAI**, is based on an ensemble of six different imputation methods.This new tool will nicely enhances data analysis capabilities in proteomics research.

Substantial missing rate



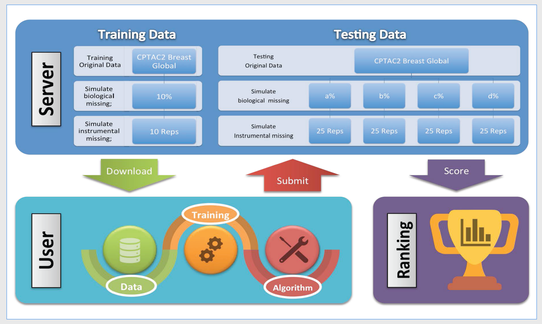
Missing not at random



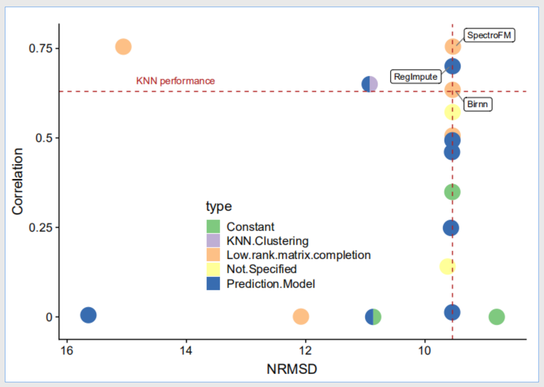
NCI-CPTAC Dream Challenge: imputation of proteomics data

The Challenge included a competition phase and a collaborative phase. In the competition phase, participants were invited to submit imputation algorithms trained on labelled LC-MS/MS proteomics data sets, and the performances of these algorithms were evaluated on a collection of test datasets generated from the CPTAC BRCA data. In the collaborative phase, together with the three winning teams from the competition phase, we further enhanced and integrated different imputation techniques and developed the final Aggregation based Imputation algorithm --- DreamAI

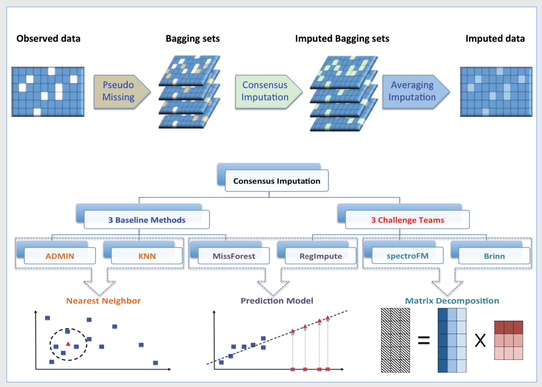
Challenge Design



Participants Performance



Structure of Dream AI



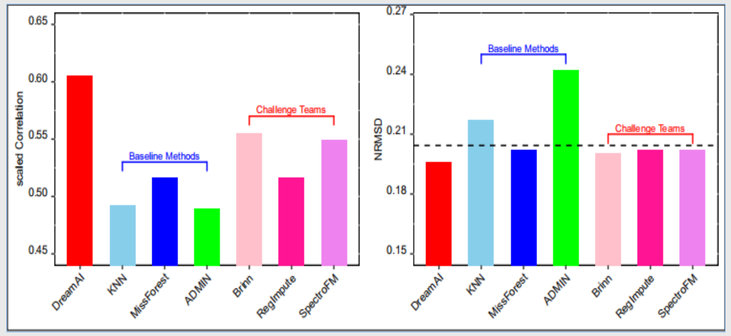
Performance of DreamAI

- DreamAI has better performance than the 6 individual algorithms on imputation of CPTAC2 OV proteomics data set (performed on samples produced by PNNL, and evaluated with the same replicated samples produced by JHU). Evaluation has been assessed by NRMSD and Correlation.

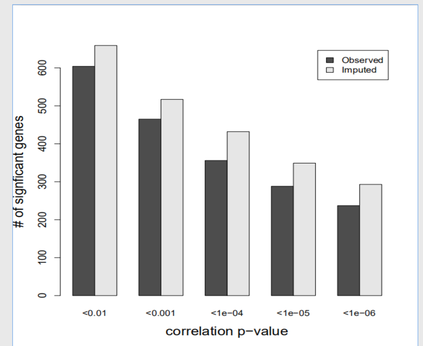
$$NRMSE = \frac{\sqrt{\sum_{i=1}^{n_{missing}} (y_i - x_i)^2 / n_{missing}}}{y_{max} - y_{min}} \qquad r = \frac{1}{n_{missing} - 1} \sum_{i=1}^{n_{missing}} \frac{(x_i - \overline{X})(y_i - \overline{Y})}{S_x S_y}$$

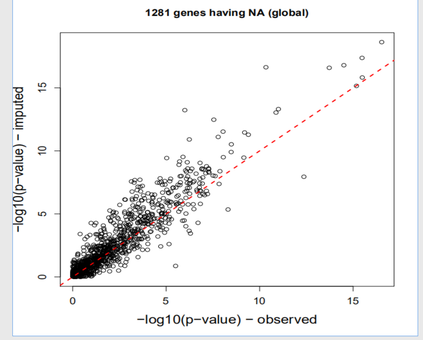
- DreamAI Imputation increased protein-rna concordance at gene level in CPTAC3 CCRCC proteomics data set. Concordance was estimated by the pairwise correlation between proteomics and gene expression

DreamAI out performed all single algorithm components in proteomics data of CPTAC2 OV cohort



DreamAI Imputation increases gene level Protein-RNA concordance in CPTAC3 CCRCC cohort



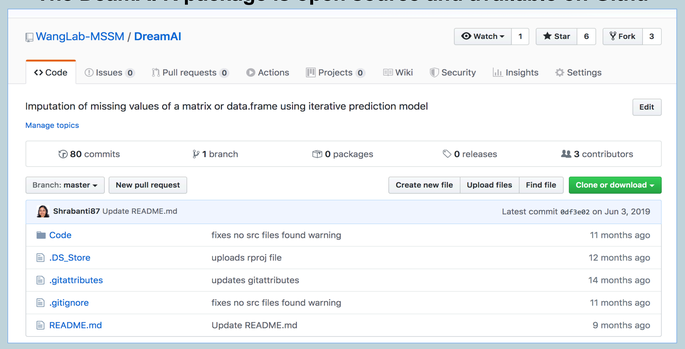


Acknowledgement and References

R Package of the algorithm is publicly available through Github:
https://github.com/WangLab-MSSM/DreamAI

Paper draft is also available on biorxiv:
https://www.biorxiv.org/content/10.1101/2020.07.21.214205v1

The DeamAI R package is open source and available on Githu



1. Mertins, Philipp, et al. "Proteogenomics connects somatic mutations to signalling in breast cancer." Nature 534, no. 7605 (2016): 55.

2. Zhang, Hui, et al. "Integrated proteogenomic characterization of human high-grade serous ovarian cancer." Cell 166, no. 3 (2016): 755-765.

3. CPTAC (NCI/NIH). “CPTAC Ovarian Cancer Confirmatory Study.” Distributed by NCI Proteomic Data Commons. https://cptac-data-portal.georgetown.edu/cptac/s/S038

4. CPTAC (NCI/NIH). “CPTAC Breast Cancer Confirmatory Study.” Distributed by NCI Proteomic Data Commons. https://cptac-data-portal.georgetown.edu/cptac/s/S039

5. Clark, David J. et al. "Integrated proteogenomic characterization of clear cell renal cell carcinoma." Cell 179, no. 4 (2019): 964-983.

6. Hastie, Trevor, Robert Tibshirani, Gavin Sherlock, Michael Eisen, Patrick Brown, and David Botstein. "Imputing missing data for gene expression arrays." (1999).

7. Stekhoven, Daniel J., and Peter Bühlmann. "MissForest—nonparametric missing value imputation for mixed-type data." Bioinformatics 28, no. 1 (2011): 112-118.

8. Wang, Minghui, et al. "The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease." Scientific data 5 (2018): 180185.