Machine Learning Prediction of Protein Abundance from Transcriptomics in CPTAC Breast Cancer

[1]Muhammad Rayyan Faher Shahab
[1]Division of Biomedical Oncology, Institute for Genetic Medicine, Hokkaido University

BACKGROUND

Protein abundance is not directly determined by transcript abundance alone because it is also influenced by translation efficiency, degradation, and post-transcriptional regulation. Multi-omics integration combined with machine learning provides a framework to model RNA-to-protein relationships at scale.

OBJECTIVE

To evaluate whether transcriptomic features can predict protein abundance using matched RNA-seq and proteomics data, and to investigate PARP1 as a biologically relevant case study.

WHY AI / MULTI-OMICS

Machine learning can capture complex relationships between transcriptomic features and protein abundance beyond simple one-gene correlation. This is relevant for understanding biological regulation and for prioritizing protein targets in large-scale omics datasets.

DATASET

CPTAC breast cancer multi-omics cohort
n = 103 tumor samples with matched:

- RNA-seq
- proteomics

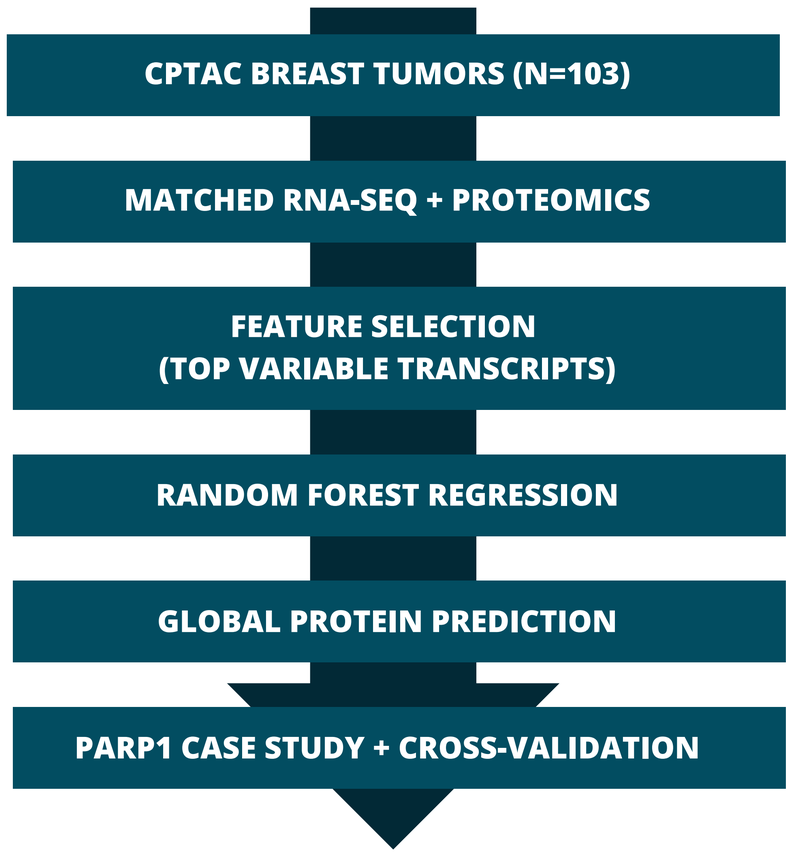
A global screen was performed across 301 proteins with sufficient sample coverage.

WORKFLOW

1. Load matched transcriptomic and proteomic data
2. Select highly variable transcript features
3. Train Random Forest regression models
4. Evaluate performance using R^2 and MAE
5. Perform global protein screening
6. Validate PARP1 model with 5-fold cross-validation

METHODS

Matched RNA-seq and proteomics data were analyzed using Python. Transcriptomic features were selected based on variance. Random Forest regression was used to predict protein abundance from transcriptomic profiles. Model performance was evaluated using R^2 and mean absolute error (MAE). PARP1 model stability was assessed using 5-fold cross-validation.



EXAMPLE PYTHON WORKFLOW

```
rf_model = RandomForestRegressor(
   n_estimators=200,
   random_state=42,
   n_jobs=-1
)

cv_scores = cross_val_score(
   rf_model, X_cv, y_cv,
            cv=KFold(n_splits=5,   shuffle=True,
random_state=42),
   scoring="r2"
)
```

RESULTS

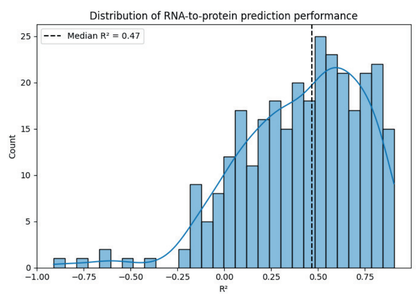


Fig 1. Distribution of prediction performance

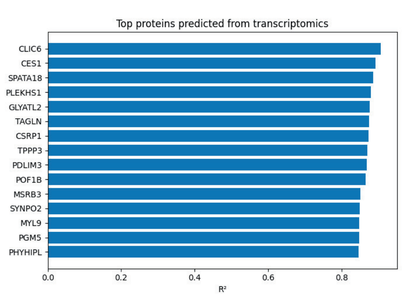


Fig 2. Top predicted proteins

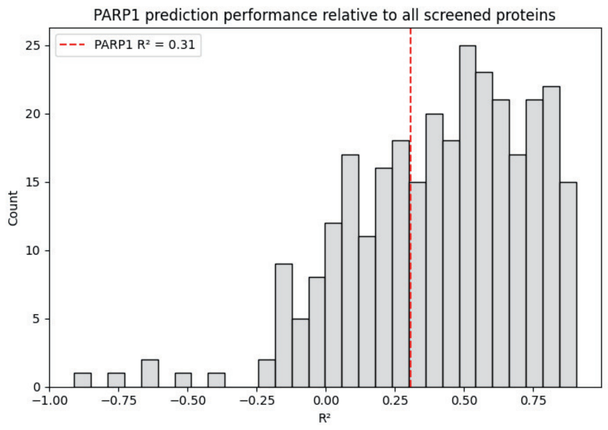


Fig 3. PARP1 prediction position

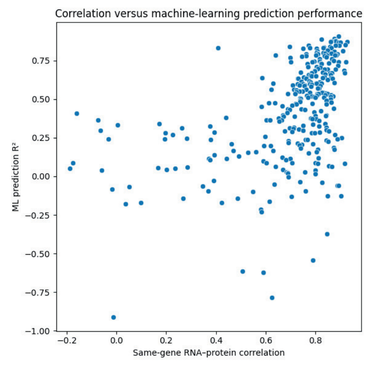


Fig 4. Correlation vs ML prediction

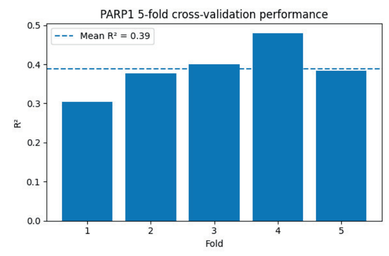


Fig. 5 PARP1 5-fold cross-validation

Gene-wise RNA–protein correlations were broadly positive, indicating overall coupling between transcriptome and proteome. Across 301 screened proteins, machine learning prediction performance was heterogeneous, with a **median R^2 of 0.47.**
For the PARP1 case study, Random Forest regression achieved:

- **Test-set R^2 = 0.31**
- **MAE = 0.22**

5-fold cross-validation showed stable performance:

- **Mean R^2 = 0.39**
- **SD = 0.06**

Feature importance analysis identified transcriptomic predictors associated with PARP1 protein abundance, including **MUC5AC, MASP1, and AKR1C3.**

DISCUSSION

These results suggest that transcriptomic features can explain a substantial fraction of protein abundance for many proteins, although predictability varies widely across targets. Proteins with high prediction performance may be more tightly regulated at the transcriptional level, while proteins with weaker prediction may be influenced more strongly by post-transcriptional or post-translational mechanisms. PARP1 showed moderate predictability, consistent with its regulation by multiple biological layers.

CONCLUSION

Machine learning enables meaningful prediction of protein abundance from transcriptomic data in matched multi-omics cohorts. In CPTAC breast cancer, global RNA-to-protein prediction showed moderate overall performance, while PARP1 demonstrated stable and reproducible prediction with cross-validation. This approach provides a practical framework for AI-based multi-omics integration.

LIMITATIONS & FUTURE WORK

- Analysis was performed in a single cohort
- Random Forest was used as an initial model; other models may improve performance
- Protein abundance is influenced by post-transcriptional regulation not captured by RNA alone
- Future work includes external validation and extension to other clinically relevant targets

REFERENCES

1. Mertins P, et al. Proteogenomics connects somatic mutations to signalling in breast cancer. Nature. 2016.
2. Krug K, et al. A curated resource for multi-omics data in cancer using CPTAC datasets.
3. Pedregosa F, et al. Scikit-learn: Machine Learning in Python. JMLR. 2011.
4. McKinney W. Data structures for statistical computing in Python. Proc. SciPy. 2010.
5. Hunter JD. Matplotlib: A 2D graphics environment. Comput Sci Eng. 2007.

ACKNOWLEDGEMENTS

This work was prepared as part of the CLAP AI course presentation.

HOKKAIDO UNIVERSITY

Institute for Genetic Medicine
HOKKAIDO UNIVERSITY