Predictive Modelling of Post-Translational Regulation During Human Cell Cycle Progression

Gregory M. Parkes, Mahesan Niranjan and Robert Ewing

Highfield Campus, University Road, Southampton, SO17 1BJ (023 8059 5000)

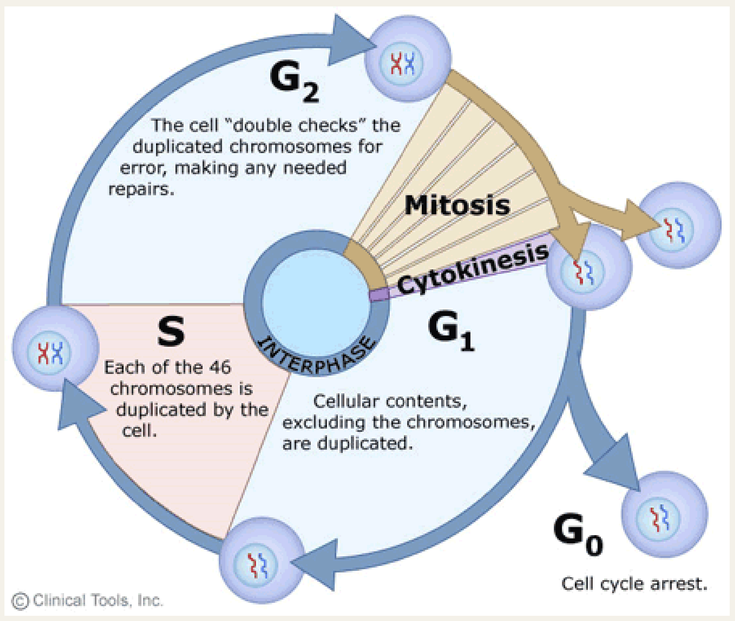
Abstract

The cell cycle is a complex phenomena of RNA and protein interactions which act as the 'kernel' of cell behaviour.

- Characterising protein interactions is crucial in understanding how cells fail (cancer, disease).
- Protein concentrations within a cell (in time) can be predicted using linear and non-linear models using corresponding mRNA and translation measurements as input.
- Outliers to the model generated are shown to contain post-translational regulatory functions by GO analysis.

Introduction

Multi-'omics (DNA, RNA, protein, epigenetic) approaches are increasingly adopted in molecular cellular analyses [1]. The central system impacting this is the cell cycle; leading in cell growth and replication.

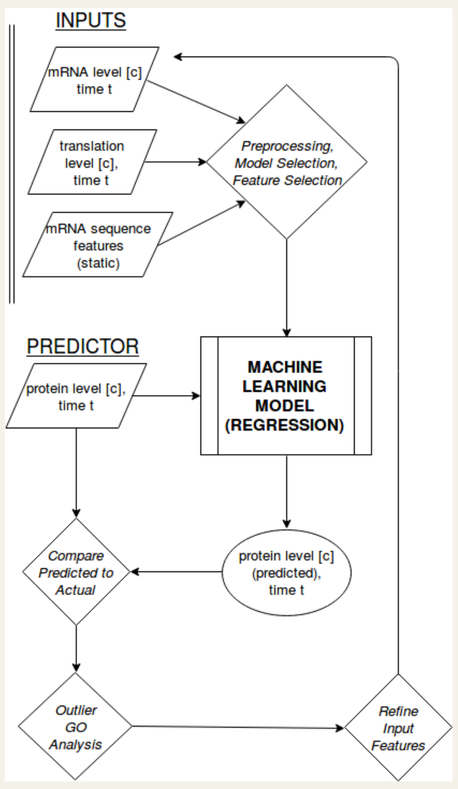


Previous models of protein concentration include both deterministic (regression)[2] and probabilistic (Bayesian, coupled-mixture)[3] approaches. These approaches have varied in success (R^2=0.5-0.86), are steady-state static samples and performed on non-human cell lines.

Methods

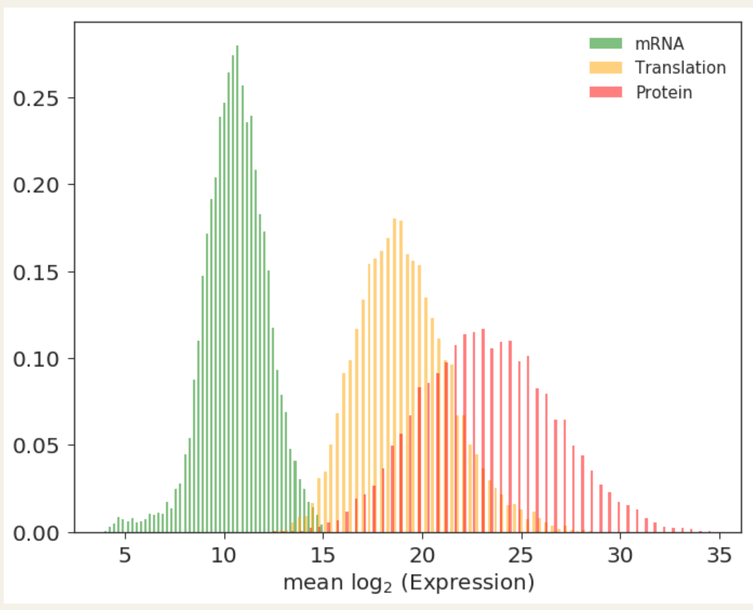
mRNA/ translation levels are drawn experimentally, sequence features are derived from online databases.

Gradient-boost tree algorithms work best across all time steps.



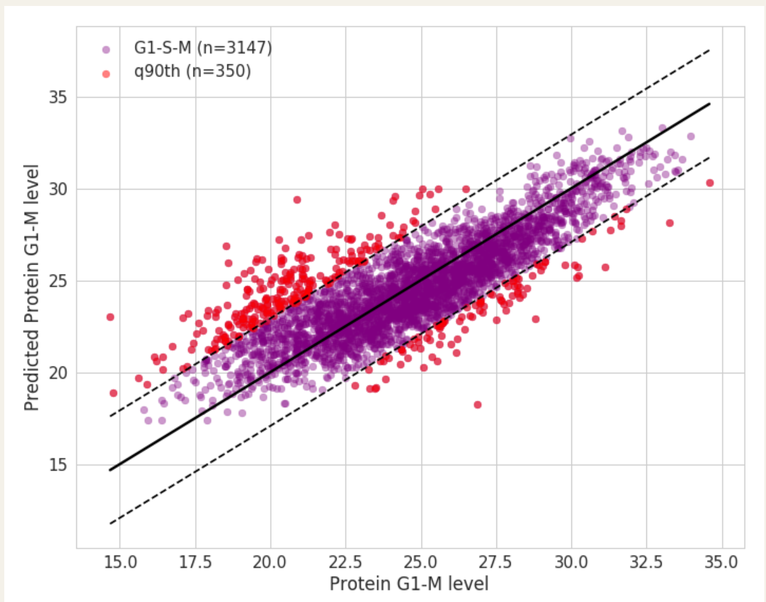
Results

We build from previous experimental work [1] which measures mRNA (microarray), translation (PUNCH-P) and protein (MS) in time-series:

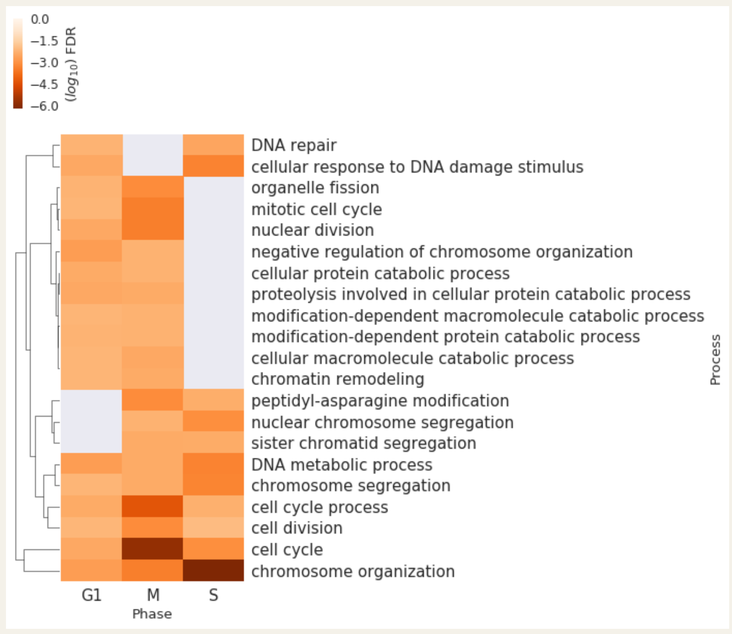


Dataset consists of 6700 RNA, 4000 translation and 5500 protein. Protein variance and subsequent complexity from a single mRNA strand leads to many-many relationships (difficult mapping). Modest correlation (r_s=0.4) between mRNA and protein less than originally anticipated.

Incorporating sequence feautres in addition as inputs to the regression model, correlation/accuracy improves substantially (r_s=0.82) between predicted and actual protein across all timesteps:



Outliers to model from *a priori* are predicted to be post-translationally regulated. We see which gene functions are *most frequent* using clustered GO analysis on outlier samples.



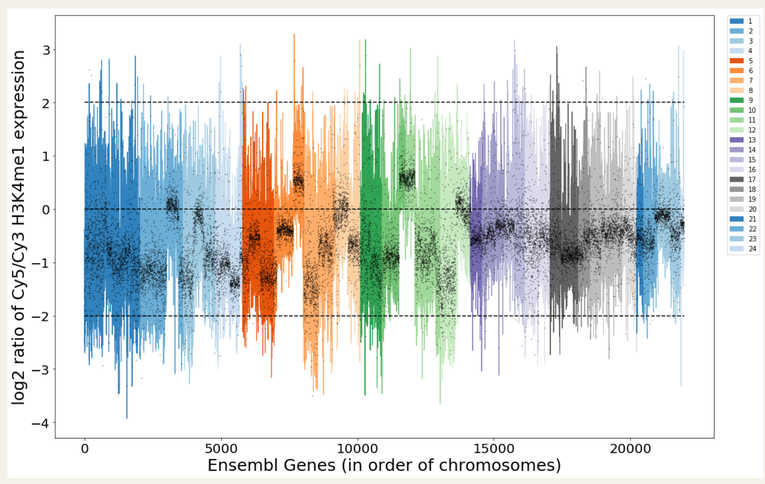
Clustering of cell cycle, protein breakdown (catabolism) and DNA-interacting processes dominate outlier function, which supports *a priori* hypothesis.

Conclusions

- Translation is vastly more important in predicting protein than mRNA (alone).
- Sequence-derived features regarding the mRNA/protein partly and cumulatively contribute somewhat to protein prediction.
- Post-translational regulation is established as dominant in outliers to a pre-protein input model – particularly overestimated proteins.
- Newly formed dataset provides a benchmark for human time-series predictive proteomics and translatomics.

Future Work

Sequence-derived features alone will not build a substantial model; thus our future work is primarily to extract additional *multi-'omic* features such as DNA methylation, halflives and histone modifications to use as input.



Extending the resolution of the time-series model will be challenging (only 3 timesteps), by integrating protein-protein interaction (PPI) networks and higher-resolution mNRA experimental data to develop confidence intervals.

Acknowledgements

The reviewing work of colleagues Omar Shetta, Pratheeba Jeyananthan and Tristian Millington is gratefully acknowleged.

Contact Information

Gregory M. Parkes

University Road,
Highfield
Campus,
Southampton,
SO17 1BJ

Email:
g.m.parkes@soton.ac.uk
Web:
cmg.soton.ac.uk/people/gmp1u16/

This work was supported by an EPSRC Doctoral Training Centre grant (EP/L015382/1)