The Influence of DNA Sequence-Derived Features across the ‘omics scales

Parkes, G.M.[1], Niranjan, M.[2]

[1]School of Electronics and Computer Science, Faculty of Engineering and Physical Sciences, University of Southampton, UK.

[2]School of Electronics and Computer Science, Faculty of Engineering and Physical Sciences, University of Southampton, UK.

Abstract

Effective modelling across the genomic scales within a cellular environment plays a crucial role in understanding the principles that govern cell cycle aberration, for instance cancer or disease. The selection of alleles, in conjunction with RNA and protein concentrations, with epigenetic factors; contribute significantly to the cell state and capacity to function. Further to this, sequence-derived features (SDFs) derived from DNA, RNA and protein sequences can contribute useful static information in conjunction with these dynamic processes to improve inference and control for steady-state effects in measurement data. These are commonly applied in transcriptomic studies whereby mRNA level acts as a proxy for protein abundance, as SDFs can be added to the model to improve predictive power. A major limiting factor of many previous studies has been lack of supportive data to coincide expression levels in the analysis of various biological domains.

Previous work has explored the relationship between mRNA expression with sequence-derived features (SDFs) against protein abundance in *S. cerevisiae* models [1, 2], from which we explored the cell cycle using an integrative approach including translation rate [3]. We found that even using a multi-’omics expression approach (mRNA, translation) without SDFs to model protein abundance produced correlations lower than expected (r=0.67) in a landmark cell cycle study [4]. We found that including an additional 15 SDFs significantly improved corrected correlations against protein abundance (see *Figure 1, r=0.83*).

Here, we present a data-driven holistic model which integrates over 180 different sequence-derived features (SDFs) across RNA and protein sequences including selective codon bias, gene length and biophysical properties with majority coverage over the curated *H. sapiens* genome. We contrast these features against multiple mRNA, translation and protein level datasets to find modest correlation between SDF-models and protein

level (r=~0.62) or mRNA level (r=~0.53). We explore which SDFs are most significant for predicting each 'omic datatype by partial correlation analysis, eliminating redundant SDFs and highlight the impact of using many small-impact factors to cumulatively contribute to analysing global expression (see *Figure 2*). In addition, mRNA or proteins found as outliers to our models invariably bare significant dynamical properties both across the cell cycle and in disease cases, thus helping to select for genes of interest. In this study, we not only explore the association between SDFs but perform predictive analysis on a number of multi-'omic datasets in different cell lines, with SDFs considered singularly for prediction but also in numerous combinations with concordant expression levels.

The expanse of SDFs (with labels) are fully available for plug-and-play into any RNA-protein and/or other 'omic dataset whereby they provide features that will control for sequence-static effects.

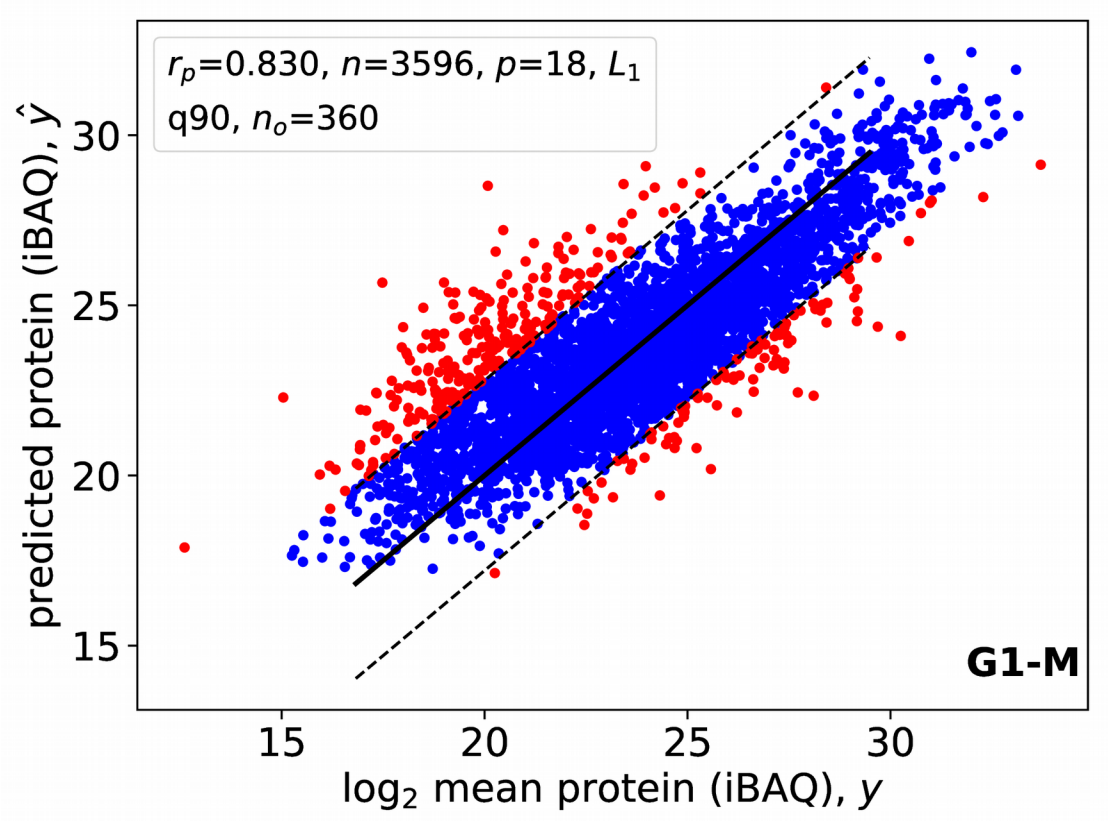

Figure 1 *Scatterplots of measured protein abundance (y) against L_1-regularized predicted protein abundance (y hat) using a gradient-boosted regression tree model and leave-one-out cross validation across the cell cycle (G1,...,M). Perfect prediction on black line (solid). 90th percentile outliers (with dashed line) w.r.t the squared-error in red. r_p is pearsons' product-moment correlation, p is the number*

of features, N_o refers to the number of outliers.

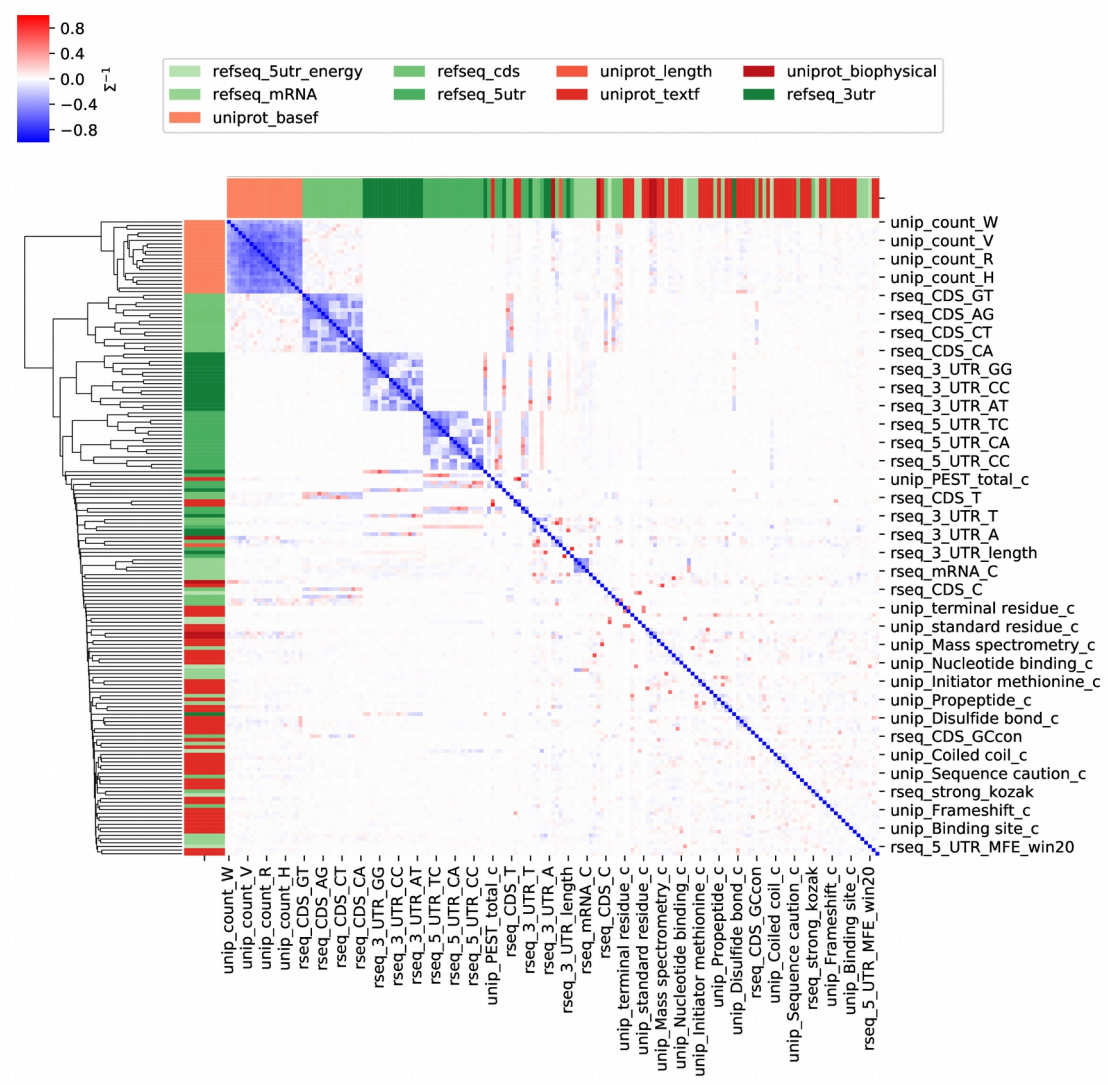

Figure 2 *Hierarchical clustering of partial correlations between sequence-derived features (SDFs), whereby each partial correlation is the correlation between column and row, eliminating every other SDF. Values furthest from zero indicate co-dependence. Green grouping bar labels refer to (R)ef(seq)/mRNA/RNA sequence features, Red/orange grouping bar labels refer to (Unip)rot/amino-acid sequence features.*

References

[1] Tuller, T., Kupiec, M. and Ruppin, E. (2007). Determinants Of Protein Abundance And Translation Efficiency In S. Cerevisiae. PLoS Computational Biology 3.12: e248.

[2] Gunawardana, Y. and Niranjan, M. (2013). Bridging the gap between transcriptome and proteome measurements identifies post-translationally regulated genes. Bioinformatics, 29(23), pp.3060-3066.

[3] Parkes, G.M. and Niranjan, M. (2019). Uncovering Extensive Post-Translation Regulation During Human Cell Cycle Progression By Integrative Multi'-omics Analysis. *BMC Bioinformatics*. (Under Review).
[4] Aviner, R., Shenoy, A., Elroy-Stein, O. and Geiger, T. (2015). Uncovering Hidden Layers of Cell Cycle Regulation through Integrative Multi-omic Analysis. PLOS Genetics, 11(10), p.e1005554.