BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 23 2013, pages 3060–3066
doi:10.1093/bioinformatics/btt537

Systems biology

Advance Access publication September 16, 2013

Bridging the gap between transcriptome and proteome measurements identifies post-translationally regulated genes

Yawwani Gunawardana and Mahesan Niranjan*

School of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, UK

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Despite much dynamical cellular behaviour being achieved by accurate regulation of protein concentrations, messenger RNA abundances, measured by microarray technology, and more recently by deep sequencing techniques, are widely used as proxies for protein measurements. Although for some species and under some conditions, there is good correlation between transcriptome and proteome level measurements, such correlation is by no means universal due to post-transcriptional and post-translational regulation, both of which are highly prevalent in cells. Here, we seek to develop a data-driven machine learning approach to bridging the gap between these two levels of high-throughput *omic* measurements on *Saccharomyces cerevisiae* and deploy the model in a novel way to uncover mRNA-protein pairs that are candidates for post-translational regulation.

Results: The application of feature selection by sparsity inducing regression (l_1 norm regularization) leads to a stable set of features: i.e. mRNA, ribosomal occupancy, ribosome density, tRNA adaptation index and codon bias while achieving a feature reduction from 37 to 5. A linear predictor used with these features is capable of predicting protein concentrations fairly accurately ($R^2 = 0.86$). Proteins whose concentration cannot be predicted accurately, taken as outliers with respect to the predictor, are shown to have annotation evidence of post-translational modification, significantly more than random subsets of similar size $P < 0.02$. In a data mining sense, this work also shows a wider point that outliers with respect to a learning method can carry meaningful information about a problem domain.

Contact: mn@ecs.soton.ac.uk

Received on June 11, 2013; revised on September 10, 2013; accepted on September 11, 2013

1 INTRODUCTION

The analysis of high-throughput experimental data has played a dominant role in biological research over the last decade or so. Advances in instrumentation, coupled with our ability to archive and share data, have revolutionized the way one approaches biological problems, more at a systems level than at the individual component level. Terabytes of data from thousands of experiments at the transcriptome, proteome and metabolome levels are now available along with metadata corresponding to the primary scientific question. There is, however, a massive skew in the amount of interest shown across the above *omic* scales, gene expression measurements made with microarray technology being highly dominant with respect to the other two. The rapid take-up of this technology by the experimental community, the monotonic reduction in cost of arrays and the early establishment of data archiving initiatives (Brazma *et al*., 2001) have led to a large community-wide focus on the transcriptome. Functional inference about co-regulated genes or genes along a signalling pathway (Brown *et al*., 2000), disease state classification focusing at the molecular level subtypes (Golub *et al*., 1999), subspace projections (Zheng-Bradley *et al*., 2010) and the reconstruction of regulatory networks (Liao *et al*., 2003; Sanguinetti *et al*., 2006) have been a number of notable success stories with transcriptome-level studies.

However, the transcriptome itself can, at best, give an approximate picture of cellular state and function. Useful biological phenomena such as dynamic cellular function and differential spatio–temporal behaviours arise from quantitatively and precisely regulating protein levels. Such behaviours arising from protein-level regulations have been modelled extensively by mathematical and computational models. Examples include controlled progression through the cell cycle (Chen *et al*., 2004), transcription delay-driven oscillations (Monk, 2003) and spatial selectivity in morphogenesis (Houchmandzadeh *et al*., 2002; Liu and Niranjan, 2011).

Several authors have evaluated the correlation between mRNA measurements and the corresponding protein measurements (Beyer *et al*., 2004; Futcher *et al*., 1999; Gygi *et al*., 1999; Wu *et al*., 2008) and report varying levels of correlation. Tuller *et al*. (2007) have developed a machine learning-based predictor of protein concentrations, which takes a different approach to previous research. In addition to mRNA levels, they construct a dataset with several properties of mRNA–protein pairs and train a linear predictor to predict protein levels. They carry out a greedy feature selection procedure to select a subset of relevant features. By this process, Tuller *et al*. (2007) achieved a correlation of 0.76 between the true concentrations and the corresponding linear predictions. Their greedy feature selection approach selects three input features as relevant predictors: (i) mRNA levels; (ii) tRNA adaptation index (tAI); and (iii) evolutionary rate (ER), determined by rate of evolution of a gene by comparison with orthologous in other organisms.

Data-driven models have been used extensively in the analysis of genomic data. Clustering, classification and time series analysis of microarray data have been carried out by several authors. Probabilistic approaches such as coupled mixture model with clustering (Rogers *et al*., 2008) and Bayesian model (Kannan *et al*., 2007) on transcriptomic and proteomic expressions investigate the relationship between these measurements. An approach that has not attracted much usage in genomic data

*To whom correspondence should be addressed.

analysis is novelty detection, in which one builds a statistical model of normal data and tests these against newly arriving abnormal data. The basic premise in such an approach is that when a data-driven model is applied to data, examples (or subsets of data) on which the model fails will also be informative. We build on this notion and, by seeking to develop a predictor of protein concentrations in the same spirit as in previous work (Tuller *et al.*, 2007), identify mRNA–protein pairs that are novel with respect to the performance of such a predictor.

We construct a data-driven linear predictor of protein concentrations, using as input mRNA concentrations and other proxy variables that can potentially regulate protein levels. Once we construct such a predictor, we look for systematic errors made by the predictor; i.e. we hypothesize that those mRNA–protein pairs for which construction of a data-driven predictor is difficult and also predicted protein abundance is lower than the measured abundance, are likely candidates for post-translational regulation. This follows from the fact that the input features used in constructing a regressor have no information pertaining to post-translational modifications (PTMs).

Post-translational regulation of proteins is important in many biological processes. For example, Tebaldi *et al.* (2012) demonstrate significant response variations at the translational level, decoupled from the transcriptional level, of mammalian cells under various stimuli. O'Neill *et al.* (2011) show that animal and plant cells have prominent post-translational contributions to timekeeping with respect to biochemical oscillations. Further, powerful computational models are also being applied to correcting measurements of post-translationally modified proteins (Chung *et al.*, 2013).

PTMs are known to be triggers of intracellular proteolytic degradation (Callis, 1995). *In vivo* stability of proteins can be substantially influenced by specific amino acid substitutions. PTMs such as phosphorylation and acetylation can act as proxies for such mutations by attachments at specific local sites, increasing the susceptibility of the protein to proteinase action (Holzer and Heinrich, 1980; Hood *et al.*, 1977). Localized PTMs, such as methylation, can be equivalent to site-specific amino acid substitutions, affecting the degradation rate of proteins (Stadtman, 1990). Nalivaeva and Turner, (2001), reviewing PTMs, suggest that glycosylation (glycoprotein) and N-link acetylation influence protein stability. They also claim modifications caused by isopeptide bond formations with members of the ubiquitine family can be implicated in protein turnover, post-translationally. Swaney *et al.*'s (2013) study shows that phosphorylation machinery can be regulated by ubiquitination.

Further, motif information on determinants of protein stability and degradation under PTMs is often available. The presence of PEST motif sequences located in flexible regions accelerates degradation under phosphorylation (García-Alai *et al.*, 2006; Marchal *et al.*, 1998). N-terminus segments act as degradation signals in cellular proteins. Thus, *N*-actelylation with *N*-acetyltransferase segments is directly involved in protein degradation process (Hwang *et al.*, 2010; Solomon and Goldberg, 1998). D and KEN Box motifs signal the anaphase promoting complex machinery that leads to ubiquitination and subsequent protein degradation (Burton and Solomon, 2001; Pfleger and Kirschner, 2000). The previously mentioned are observations we will exploit to confirm that proteins found by our novelty-detection framework are likely candidates for post-translational regulation of their concentrations (see Section 3).

This article makes two contributions to data-driven modelling at the transcriptome–proteome interface. First, the linear regression with sparsity inducing regularization (LASSO) method can identify features that are relevant to a prediction problem. This, in the context of computational biology problems, is an alternate approach to the often used greedy forward selection of features. The accuracy of prediction of protein concentrations shows improvement over previous efforts at this problem. Second, model failures carry useful information, and this is demonstrated by identifying genes whose predicted protein concentrations are outliers (Li and Niranjan, 2006) with respect to predictions obtained by a global regression. These are confirmed by checking functional annotations.

2 METHODS

2.1 Data preparation

Several datasets were combined together using the open reading frame (ORF) and gene names to generate our final dataset. mRNA abundance data for *Saccharomyces cerevisiae* were downloaded from Greenbaum *et al.* (2003). We used PaxDb (Wang *et al.*, 2012) to find the relevant protein abundance data, which was developed by integrating four datasets (de Godoy *et al.*, 2008; Desiere *et al.*, 2006; Ghaemmaghami *et al.*, 2003; Newman *et al.*, 2006). Ribosome density was taken from Arava *et al.* (2003). Gene length, ribosomal occupancy, proteins per second and relative translation rate data were obtained from Greenbaum *et al.* (2003). mRNA half-life data were downloaded from Miller *et al.* (2011). Twenty-eight sequence-derived properties, also used by Tuller *et al.* (2007), were obtained from Cherry *et al.* (2012). tAI data were taken from Man and Pilpel (2007) and ERs of proteins were downloaded from Wall *et al.* (2005). In all cases, experimental data used corresponded to *S.Cerevisiae* cell cultures under exponential growth conditions. Comparing with previous work (Tuller *et al.*, 2007), gene length, ribosomal occupancy, proteins per second, ribosome density, relative translation rate and mRNA half-live are used as new features in our study. When these different datasets are put together, and some data are filtered for missing values and low mRNA abundances (log expression of -1), we obtained feature values and protein concentrations for 1895 proteins, which was the dataset we worked with.

2.2 Sparse regression

Feature selection is a key step in regression problems. For technical reasons, it is usually beneficial to reduce the dimensionality of the space, thereby avoiding the *curse of dimensionality*, which states that the amount of data needed to reliably estimate probability densities grows exponentially with dimensions. Further, by selecting a subset of features, we are likely to improve our ability to explain useful aspects of the problem domain. The search for a subset of features has combinatorial complexity and greedy searches such as sequential forward selection and backward deletion are commonly used (Lovell *et al.*, 1998). For the protein concentration prediction problem, Tuller *et al.* (2007) used greedy forward selection. This approach is particularly weak when there are correlated features in the input data. We chose the alternate approach of sparsity inducing regularizers embedded within the estimation of linear regression, also known as LASSO (Tibshirani, 1994), to achieve feature selection. This l_1-regularized regression has attracted much interest in recent literature and has the appealing property of easy implementation via convex

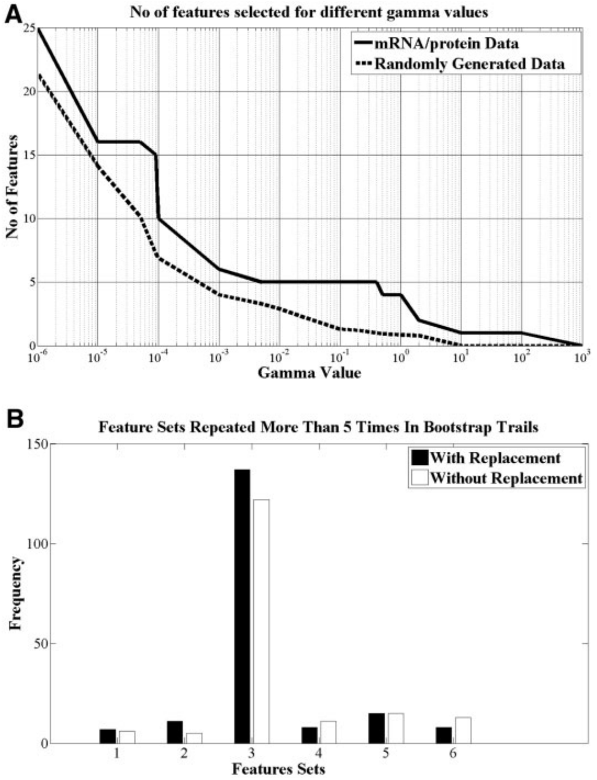

Fig. 1. Feature selection by l_1 regularization. (**A**) Variation in the average number of selected features as a function of the regularization parameter γ, which have a stable region over three orders of magnitude of γ (0.001 and 1). (**B**) Identification of the best set of features (set 3) from the most frequent six sets of features, which recognized from the stable γ region

programming (Lu *et al.*, 2011; Park and Casella, 2008; Wu *et al.*, 2009). The objective function minimized is as follows:

$$\min\{||\mathbf{X}\,\mathbf{w} - \mathbf{y}||^2 + \gamma||\mathbf{w}||_1\} \tag{1}$$

where X is the input matrix of covariates, y is the response vector and w is the weight vector of unknowns. γ controls the amount of regularization, and with the l_1 norm constraint determines the number of non-zero terms in w, i.e. sparsity of the solution. We used the `CVX` package within a `MATLAB` environment for optimization of the sparse regressor and, after observing a histogram of the resulting weight values, centre clipped the weights at 0.2 to arrive at the sparse solution.

To evaluate uncertainties in estimates, we constructed 1000 bootstrapped samples of 500 genes each from the data and estimated the sparse regressor over 20 values of γ in the range 0–1000. Average number of features selected (Fig. 1A) shows a stable region over several orders of magnitude of γ, from which a stable feature set is selected.

2.3 Development of protein abundance predictor

The protein abundance predictor is a linear predictor, based on the five features selected by the LASSO method, obtained by minimizing the following:

$$\min\{||\mathbf{X}\,\mathbf{w} - \mathbf{y}||^2\} \tag{2}$$

Data were partitioned into five groups at random. With each of the groups retained as test data, linear models were estimated from the remaining four groups pooled together. Thus, all predicted values of protein concentrations from which outliers were detected (see Section 3) were on out-of-sample predictions. Predictors were developed with the five features selected by LASSO, the set of three features from previous work (Tuller *et al.*, 2007) and with all 37 features as input.

Neural net: We also implemented neural network predictors to confirm any non-linear relationships between the variables and output protein concentrations. For this the neural network toolbox in `MATLAB` was used with stochastic gradient descent optimization of a multi-layer perceptron neural network with 10 hidden units (Bishop, 1995).

2.4 PTM annotation check

We looked for outliers being post-translationally regulated by observing the functional annotations at two levels. At the first level, we used `UniProt` database (Magrane and Consortium, 2011), which is cross-referred by the PaXDb (Wang *et al.*, 2012) where we obtained our initial protein abundances. Several databases were used to carry out the finer level annotation check. EMBOSS explorer `epestfind` database (Rice *et al.*, 2000) was used to detect PEST motifs of the proteins with phosphorylation modification. *N*-termini segments of acetylation were obtained by `NetAcet 1.0` database (Kiemer *et al.*, 2005). D and KEN box motifs, which accelerate ubiquitination, were detected using `GPS-ARM 1.0` toolkit (Liu *et al.*, 2012).

3 RESULTS

Sparse linear regression selects a compact set of features relevant for predicting protein concentrations accurately. Further, outliers with respect to the predictor we constructed, for whom the predicted protein concentration ($\hat{P}$) was greater than the measurement (P), contained significant over-representation for proteins annotated with keywords of PTMs.

3.1 Feature selection

In implementing l_1-regularized regression, the choice of regularization term γ is crucial. Figure 1(A) shows the variation in average number of retained features, as a function of γ. We note that the number of features selected does not reduce linearly. Instead, there is a stable region, over three orders of magnitude of γ (0.001 and 1) in which five features are selected, suggesting that this dataset consist of five dominant features. To confirm this, we constructed several datasets of similar size with uniform random numbers and carried out such sparse regressions. We found the monotonic reduction in the number of features selected, also shown in Figure 1A (dashed line), on the random regression problems had no stable region of a constant number of features retained.

Across the 1000 bootstrap samples (see Section 2), six combinations of feature subsets (containing 6, 5, 5, 6, 4 and 4 features) were frequently identified as relevant for the prediction. The corresponding frequencies are shown in Figure 1B. We note that feature set identified as set 3, consisting of five features, appears significantly more number of times than any of the others. This set consisted of the following features: mRNA, ribosomal occupancy, ribosome density, tAI and codon bias. This differs slightly from Tuller *et al.*'s (2007) study that identified mRNA, tAI and ER as relevant features. We find that in addition to mRNA and sequence derived features, measurements relating to translational efficiency (ribosomal density and occupancy) are also significant. This is to be expected because translation efficiency directly influences the quantity of protein synthesized (Greenbaum *et al.*, 2003).

Codon bias refers to differences in the frequency of occurrences of synonymous codons in coding DNA and evolutionary origins of codon bias has been investigated by Wallace *et al.* (2013). The role of such evolutionarily accrued biases in encoding on protein concentrations has been noted previously (Brockmann *et al.*, 2007; Tuller *et al.*, 2010). tAI might have a similar role, being related to the codon adaption index for a gene (Reis *et al.*, 2004). In a study on the human proteome, Waldman *et al.* (2010) directly associate tAI with translational efficiency.

3.2 Protein abundance predictor

With the five features we selected by l_1-regularized regression, protein abundances were predictable to a higher level of accuracy by a linear predictor than either simply looking for correlation with mRNA levels or by the features identified in Tuller *et al.*'s (2007) work. Our best five features gave a correlation of $r^2 = 0.86$ between predicted and true values, whereas the combination of mRNA, tAI and ER, identified in Tuller *et al.*'s (2007) work, gave only $r^2 = 0.80$. Using all 37 features also achieved $r^2 = 0.80$ on unseen (cross-validated) data, which is lower than our five feature accuracy. Thus, in various combinations of tests the five features selected from regularization turned out to be superior.

Performing prediction on neural net non-linear model gave only $r^2 = 0.82$ for our feature set and $r^2 = 0.79$ for the three feature combination from previous work. When the neural net was trained on all 37 features, the accuracy of prediction dropped drastically to $r^2 = 0.69$. Thus, similar to the observation made by Tuller *et al.* (2007), there is no significant advantage in using a non-linear model to this prediction task.

As ER was not selected as a dominant feature in our feature selection model, we examined the prediction performances by progressively adding our five features and then including ER as a sixth feature. As shown in Figure 2, the inclusion of the five features monotonically improved prediction results (this happens to be true for any order in which they are taken), but when ER was taken as an additional feature, the results dropped to $r^2 = 0.80$. ER as the only feature achieves a correlation of $r^2 = -0.46$ with protein concentrations. Similar to adaptation indices, the role of ER as a predictor of protein concentration is merely an empirical observation noted by researchers (e.g. Moreira *et al.*, 2002), but the precise molecular mechanism of regulation remains unknown.

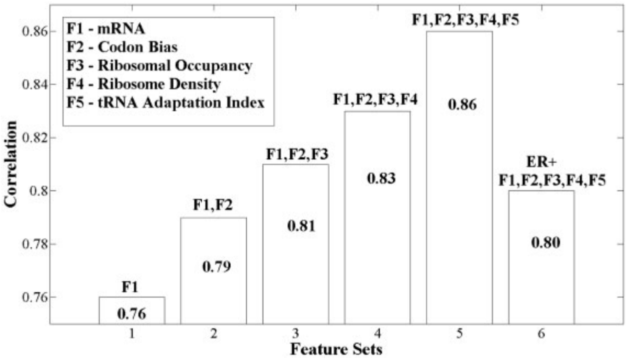

Fig. 2. Performance of l_1-regularized regressor, adding features one at a time. Addition of our mRNA, codon bias, ribosomal occcupancy, ribosome density and tAI, which identified by l_1-regularization feature selection process, monotonically increased the accuracy in each step. However, addition of evolutionary rate reduced the overall accuracy of the predictor

3.3 Post-translational regulations

Figure 3 shows a scatter plot of the predicted protein concentration ($\hat{P}$) against the true concentration (P) from which we detected outliers, points that are furthest away from the regression line (shown as solid line). When we select the top 50 outliers, 48 of them were found to be in the upper half of the graph where $P < \hat{P}$, i.e. the measured concentration is smaller than what the global regression predicts from mRNA level information.

To confirm that proteins for which $P < \hat{P}$ are likely candidates for post-translational regulation, we carried out an analysis using functional annotations at two levels: (i) at a coarse level, PTMs are a primary requirement for regulation and (ii) at a finer level, PTMs coupled with information about protein stability determinants (motifs) are stronger indicators of post-translational regulation (i.e. Phosphorylation + PEST motifs, Acetylation + N-termini segments and Ubiquitination + D or KEN Box motifs). At both levels, we looked for over-representation of annotations within the outlier set when compared with random subset of same size.

3.4 Level 1: coarse level PTM analysis

Forty-two proteins among the 48 outliers (upper half) were recognized as being subject to PTMs and are shown in Table 1. Neither of the proteins found as outliers in the lower half of the scatter plot had this property.

To estimate a level of confidence in the PTM keyword over-representation in the outlier set, we used 1000 random samples of proteins of size 50 and constructed a Gaussian distribution of the number of PTM proteins found in these sets. The resulting distribution had mean and standard deviation of 34.286 and 3.576, respectively. From this the claim of over-representation of PTM proteins among the outlier subset can be made at significance of $P = 0.02$. As is usual in biomedical research of this kind (McDonald, 2009), if we take a $P = 0.05$ as a threshold of

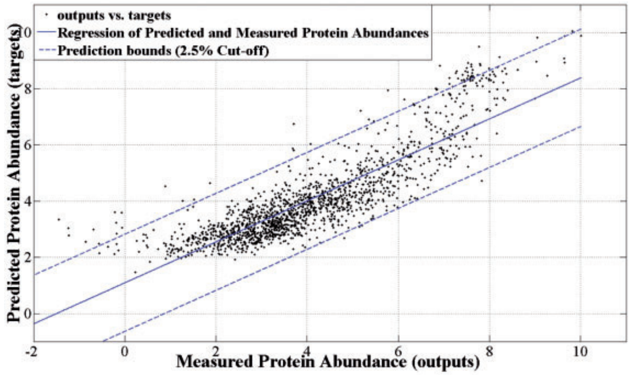

Fig. 3. Outlier detection on protein concentration prediction. In all, 2.5% of the least accurate predictions (a total of 50) were selected as outliers

Table 1. PTMs identified in 50 outliers (cut-off at 2.5%)

ORF name	Gene name	PTMs
YJL129C	TRK1	Glycoprotein, Phosphoprotein
YBR038W	CHS2	Glycoprotein, Phosphoprotein
YDL093W	PMT5	Glycoprotein
YDL217C	TIM22	x
YFL029C	CAK1	Phosphoprotein
YHR031C	RRM3	Phosphoprotein
YJR124C	YJR124C	Phosphoprotein
YDL048C	STP4	Phosphoprotein
YGL159W	YGL159W	x
YDR006C	SOK1	Phosphoprotein
YIL169C	YIL169C	Glycoprotein
YDL222C	FMP45	Glycoprotein, Phosphoprotein
YDL130W	RPP1B	Acetylation, Phosphoprotein
YCR010C	ADY2	Phosphoprotein
YHR141C	RPL42B	Methylation
YBR106W	PHO88	Phosphoprotein
YAR075W	YAR075W	Phosphoprotein
YHR094C	HXT1	Glycoprotein, Phosphoprotein
YDR342C	HXT7	Glycoprotein, Isopeptide b., Phosphoprotein, Ubl con.
YBR1317	RPS9B	Phosphoprotein
YJL177W	RPL17B	Phosphoprotein
YGR282C	BGL2	Glycoprotein
YBL0613	RPS8A	Phosphoprotein
YDR225W	HTA1	Acetylation, Isopeptide b., Phosphoprotein, Ubl conj.
YEL027W	VMA3	x
YKR059W	TIF1	Acetylation, Phosphoprotein
YGL030W	YGL030W	Phosphoprotein
YIL148W	RPL40A	Isopeptide b., Phosphoprotein, Ubl con.
YBR010W	HHT1	Acetylation, Methylation, Phosphoprotein
YHR021C	RPS27B	Phosphoprotein
YGR034W	RPL26B	Phosphoprotein
YER102W	RPS8B	Phosphoprotein
YDL083C	RPS16B	Acetylation, Phosphoprotein
YDR064W	RPS13	Phosphoprotein
YCR031C	RPS14A	Acetylation, Phosphoprotein
YDL081C	RPP1A	Acetylation, Phosphoprotein
YEL034W	HYP2	Acetylation, Phosphoprotein
YDR447C	RPS17B	Phosphoprotein
YER117W	RPL23B	Acetylation, Methylation, Phosphoprotein
YKL180W	RPL17A	Phosphoprotein
YKL056C	TMA19	x
YKL152C	GPM1	Phosphoprotein
YLR044C	PDC1	Acetylation, Phosphoprotein
YCR012W	PGK1	Acetylation, Phosphoprotein
YGL123W	RPS2	Acetylation, Phosphoprotein
YDR382W	RPP2B	Phosphoprotein
YGR148C	RPL24B	Phosphoprotein
YDL014W	NOP1	Methylation, Phosphoprotein
YDL080C	THI3	x (lower outlier region)
YER070W	RNR1	x (lower outlier region)

Note: Ubl con. stands for Ubl conjugation and Isopeptide b. stands for Isopeptide bond.

Table 2. Confidence levels indicating how well the outlier subset identifies post-translationally modified proteins, at different numbers of chosen outliers

Percentage outliers (%)	No of outliers	No of PTMs ($P < \hat{P}$)	*P*-Value
1.0	20	19	0.01
2.5	50	42	0.02
5.0	100	73	0.17

Note: 1000 random trials were used in each case.

accepting a hypothesis of interest, our suggestion that proteins in $P < \hat{P}$ outlier set are post-translationally modified is supported.

We also checked 50 outliers from the Tuller *et al.*'s (2007) three feature set predictor for significance of over-representation of PTM proteins. Thirty-seven proteins were identified with PTM annotations, giving $P = 0.22$. When we took 50 outliers directly from a scatter plot of mRNA and protein levels, the number of PTMs detected was 35, corresponding to a $P = 0.42$.

We also looked at various cut-off levels at which an mRNA–protein pair could be called an outlier with respect to the global predictor. Setting cut-offs to extract 1, 2.5 and 5% of the data as outliers, we repeated the above exercise and obtained *P*-values. These are shown in Table 2.

With our five feature predictor, the top 100 outliers containing over-represented post-translationally modified proteins are at a higher level of significance than for the top 50 outliers detected from Tuller *et al.*'s (2007) three feature predictor ($P = 0.17$ and $P = 0.22$, respectively). This further confirms that the ranking of data arising from our five input predictor is more informative.

3.5 Level 2: finer level PTM analysis

At this level of probing annotations of the outlier set of proteins, 37 of the 50 had PTM with motif information. The corresponding confidence level, computed similarly to the level 1 check, achieved $P < 10^{-12}$.

For predictors with Tuller *et al.*'s (2007) feature set and for simply considering mRNA–protein scatter plot to pick outliers, the finer level annotation that gave higher levels of confidence in over-representation ($P = 0.0017$ for 26 proteins and $P = 0.042$ for 22 proteins, respectively).

When we changed the cut-off levels to 1 (20 proteins) and 5% (100 proteins) of the data defined as outliers, we obtained confidence levels of $P = 10^{-12}$ and $P = 0.001$, respectively.

We note that this level of checking annotations information, i.e. incorporating PTMs with motif information that influence protein stability, gives higher levels of confidence in support of our hypothesis.

Thus, in all checks carried out comparing available annotation information, we can conclude that the outlier set of proteins are more likely to be regulated post-translationally. Further, PTM detection ability of our predictor (by looking at the outliers) outperformed in both annotation checks.

Gene Ontology (GO) enrichment analysis: We also subjected the 50 outliers to GO enrichment analysis using Gene Ontology Enrichment Analysis Software Toolkit (GOEAST) (Zheng and Wang, 2008). Thirty-seven GO annotations were found in the outlier set, four of which were common to >30 genes (GO:0044444, GO:0009058, GO:1901576 and GO:0032991), and were found in cellular component and biological process categories. We also observed that our outlier set is enriched for ribosomal proteins with 14 GO terms relating to the ribosome.

Role of ribosomal proteins: Ribosomal genes are known to undergo intense transcriptional activity coupled with efficient translation (Warner, 1999), followed by several PTMs such as methionine removal, N-terminal acetylation, N-terminal methylation, lysine N-methylation and phosphorylation (Carroll *et al*., 2008). As our outlier set of 50 proteins contained 23 ribosomal proteins, we evaluated the effect of the dominance of ribosomal proteins on out methodology. Though several of the ribosomal proteins in the dataset had high expression levels, their distribution was not significantly different from the remainder. We repeated the entire analysis after removing the 155 ribosomal proteins from the dataset. With the reduced set, when we took 50 outliers (3%) 42 had PTM annotations at the level 1 of our check $P = 0.02$ and $P = 36$ had PTM annotation at the level of $P < 10^{-12}$. This confirms that the dominance of ribosomal proteins did not unduly influence the methodology.

Analysis of protein half-life: We checked if the prediction ability had any systematic variability that was influenced by protein half-life, i.e. concentrations of rapidly degrading proteins likely to be under-quantified. We compared absolute and squared errors of our predictor against protein half-lives published by Belle *et al*. (2006) and found no significant correlation. Of the 50 outliers we detected, protein half-life data were available for only 26 proteins, and they showed no systematic behaviour.

3.6 Discussion and conclusion

In this work, we have shown that by constructing a machine learning based predictor of cellular protein concentrations, based on the corresponding mRNA levels and other features pertaining to transcription regulation, we can identify, as outliers, proteins that are likely candidates for post-translational regulation. Of the proteins we identify as outliers, proteins that are annotated as being subject to PTMs are significantly over-represented than in any random subsets of similar size. We will not be able to get a perfect ranking in which all post-translationally regulated genes come on top. This is because a gene being annotated as being subjected to PTM, need not be modified under all conditions, several such restrictions are condition specific and for richer experimental data will be required to give a complete picture.

Two generic points also need to be mentioned in closing. First, when fitting a data-driven model in the analysis of high-throughput data, outliers or model failures can carry useful information. Unlike previous authors who focused on the correlation between mRNA and protein levels, and on building accurate protein concentration predictors, our method, by looking at model failures, extracts potentially useful information about how these proteins may be regulated. This is an example of a wider point about the use of machine learning in computational biology; i.e. the purpose, unlike in building a voice recognition or finger-print recognition system where performance is measured in terms of accuracy of classification, in biology what we require is to cut down the space over which experimental work needs to be carried out to confirm biological function, PTMs in our case. Ultimately though, proof of biological function is confirmed in wet-laboratory experimental findings. What machine learning can offer is to find a reliable reduction in the space over which such experimental explorations need to be carried out.

Second, the dataset we put together is synthesized from several different experiments carried out by different authors in different laboratories. Though all the experiments correspond to a particular organism (*S.cerevisiae*) growing under well-defined (exponential growth) conditions, there is bound to be variability in the data resulting from the fact that the different measurements were not taken from identical laboratory conditions. It is difficult to quantify the effect of such variability in the results we report.

Conflict of Interest: none declared.

REFERENCES

Arava,Y. *et al.* (2003) Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA*, **100**, 3889–3894.

Belle,A. *et al.* (2006) Quantification of protein half-lives in the budding yeast proteome. *Proc. Natl Acad. Sci. USA*, **103**, 13004–13009.

Beyer,A. *et al.* (2004) Post-transcriptional expression regulation in the yeast *Saccharomyces cerevisiae* on a genomic scale. *Mol. Cell. Proteomics*, **3**, 1083–1092.

Bishop,C.M. (1995) *The Multi-Layer Perceptron*. Oxford University Press, UK.

Brazma,A. *et al.* (2001) Minimum information about a microarray experiment (miame) toward standards for microarray data. *Nat. Genet*., **29**, 365–371.

Brockmann,R. *et al.* (2007) Posttranscriptional expression regulation: what determines translation rates? *PLoS Comput. Biol*., **3**, e57.

Brown,M.P. *et al.* (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.

Burton,J.L. and Solomon,M.J. (2001) D box and KEN box motifs in budding yeast Hsl1p are required for APC-mediated degradation and direct binding to Cdc20p and Cdh1p. *Genes Dev*., **15**, 2381–2395.

Callis,J. (1995) Regulation of protein degradation. *Plant Cell*, **7**, 845.

Carroll,A. *et al.* (2008) Analysis of the arabidopsis cytosolic ribosome proteome provides detailed insights into its components and their post-translational modification. *Mol. Cell. Proteomics*, **7**, 347–369.

Chen,K. *et al.* (2004) Integrative analysis of cell cycle control in budding yeast. *Mol. Biol. Cell*, **15**, 3841–3862.

Cherry,J.M. *et al.* (2012) *Saccharomyces genome* database: the genomics resource of budding yeast. *Nucleic Acids Res*., **40**, D700–D705.

Chung,C. *et al.* (2013) Nonparametric bayesian approach to post-translational modification refinement of predictions from tandem mass spectrometry. *Bioinformatics*., **29**, 821–829.

de Godoy,L. *et al.* (2008) Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature*, **455**, 1251–1254.

Desiere,F. *et al.* (2006) The peptideatlas project. *Nucleic Acids Res*., **34(Suppl. 1)**, D655–D658.

Futcher,B. *et al.* (1999) A sampling of the yeast proteome. *Mol. Cell. Biol*., **19**, 7357–7368.

García-Alai,M.M. *et al.* (2006) Molecular basis for phosphorylation-dependent, pest-mediated protein turnover. *Structure*, **14**, 309–319.

Ghaemmaghami,S. *et al.* (2003) Global analysis of protein expression in yeast. *Nature*, **425**, 737–741.

Golub,T.R. *et al.* (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Greenbaum,D. *et al.* (2003) Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol*., **4**, 117.

Gygi,S. *et al.* (1999) Correlation between protein and mRNA abundance in yeast. *Mol. Cell. Biol*., **19**, 1720–1730.

Holzer,H. and Heinrich,P.C. (1980) Control of proteolysis. *Ann. Rev. Biochem*., **49**, 63–91.

Hood,W. *et al.* (1977) Increased susceptibility of carbamylated glutamate dehydrogenase to proteolysis. *Acta Biol. Med. Ger.*, **36**, 1667–1672.

Houchmandzadeh,B. *et al.* (2002) Establishment of developmental precision and proportions in the early *Drosophila* embryo. *Nature*, **415**, 798–802.

Hwang,C. *et al.* (2010) N-terminal acetylation of cellular proteins creates specific degradation signals. *Science*, **327**, 973–977.

Kannan,A. *et al.* (2007) A Bayesian model that links microarray mRNA measurements to mass spectrometry protein measurements. In: Speed,T. and Huang,H. (eds) *Research in Computational Molecular Biology*. Springer, Berlin, Heidelberg, pp. 325–338.

Kiemer,L. *et al.* (2005) NetAcet: prediction of N-terminal acetylation sites. *Bioinformatics*, **21**, 1269–1270.

Li,H. and Niranjan,M. (2006) Outlier detection in benchmark classification tasks. In: *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference*. Vol. 5, ICASSP, Toulouse, pp. V557–V560.

Liao,J.C. *et al.* (2003) Network component analysis: Reconstruction of regulatory signals in biological systems. *Proc. Natl Acad. Sci. USA*, **100**, 15522–15527.

Liu,W. and Niranjan,M. (2011) The role of regulated mRNA stability in establishing bicoid morphogen gradient in *Drosophila* embryonic development. *PLoS One*, **6**, e24896.

Liu,Z. *et al.* (2012) GPS-ARM: computational analysis of the APC/C recognition motif by predicting D-Boxes and KEN-Boxes. *PLoS One*, **7**, e34370.

Lovell,D. *et al.* (1998) Feature selection using expected attainable discrimination. *Pattern Recognit. Lett.*, **19**, 393–402.

Lu,Y. *Pattern Recognition Letters* (2011) A lasso regression model for the construction of microRNA-target regulatory networks. *Bioinformatics*, **27**, 2406–2413.

Magrane,M. and Consortium,U. (2011) Uniprot knowledgebase: a hub of integrated protein data. *Database*, **2011**, bar009.

Man,O. and Pilpel,Y. (2007) Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species. *Nat. Genet.*, **39**, 415–421.

Marchal,C. *et al.* (1998) A PEST-like sequence mediates phosphorylation and efficient ubiquitination of yeast uracil permease. *Mol. Cell. Biol.*, **18**, 314–321.

McDonald,J.H. (2009) *Basic Concepts of Hypothesis Testing*. Vol. 2. Sparky House Publishing, Baltimore, MD.

Miller,C. *et al.* (2011) Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. *Mol. Syst. Biol.*, **7**, 458–470.

Monk,N. (2003) Oscillatory expression of hes1, p53, and NF-κB driven by transcriptional time delays. *Curr. Biol.*, **13**, 1409–1413.

Moreira,D. *et al.* (2002) Evolution of eukaryotic translation elongation and termination factors: variations of evolutionary rate and genetic code deviations. *Mol. Biol. Evol.*, **19**, 189–200.

Nalivaeva,N.N. and Turner,A.J. (2001) Post-translational modifications of proteins: acetylcholinesterase as a model system. *Proteomics*, **1**, 735–747.

Newman,J. *et al.* (2006) Single-cell proteomic analysis of *S. cerevisiae* reveals the architecture of biological noise. *Nature*, **441**, 840–846.

O'Neill,J.S. *et al.* (2011) Circadian rhythms persist without transcription in a eukaryote. *Nature*, **469**, 554–558.

Park,T. and Casella,G. (2008) The bayesian lasso. *J. Am. Stat. Assoc.*, **103**, 681–686.

Pfleger,C.M. and Kirschner,M.W. (2000) The KEN box: an APC recognition signal distinct from the D box targeted by Cdh1. *Genes Dev.*, **14**, 655–665.

Reis,M. *et al.* (2004) Solving the riddle of codon usage preferences: a test for transaltional section. *Nucleic Acids Res.*, **32**, 5036–5044.

Rice,P. *et al.* (2000) EMBOSS: the European molecular biology open software suite. *Trends Genet.*, **16**, 276–277.

Rogers,S. *et al.* (2008) Investigating the correspondence between transcriptomic and proteomic expression profiles using coupled cluster models. *Bioinformatics*, **24**, 2894–2900.

Sanguinetti,G. *et al.* (2006) Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities. *Bioinformatics*, **22**, 2775–2781.

Solomon,V. *et al.* (1998) The N-end rule pathway catalyzes a major fraction of the protein degradation in skeletal muscle. *J. Biol. Chem.*, **273**, 25216–25222.

Stadtman,E. (1990) Covalent modification reactions are marking steps in protein turnover. *Biochemistry*, **29**, 6323–6331.

Swaney,D.L. *et al.* (2013) Global analysis of phosphorylation and ubiquitylation cross-talk in protein degradation. *Nat. Methods*, **10**, 676–682.

Tebaldi,T. *et al.* (2012) Widespread uncoupling between transcriptome and translatome variations after a stimulus in mammalian cells. *BMC Genomics*, **13**, 220.

Tibshirani,R. (1994) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Stat. Methodol.*, **58**, 267–288.

Tuller,T. *et al.* (2007) Determinants of protein abundance and translation efficiency in *S. cerevisiae*. *PLoS Comput. Biol.*, **3**, e248.

Tuller,T. *et al.* (2010) Translation efficiency is determined by both codon bias and folding energy. *Proc. Natl Acad. Sci. USA*, **107**, 3645–3650.

Waldman,Y.Y. *et al.* (2010) Translation efficiency in humans: tissue specificity, global optimization and differences between developmental stages. *Nucleic Acids Res.*, **38**, 2964–2974.

Wall,D.P. *et al.* (2005) Functional genomic analysis of the rates of protein evolution. *Proc. Natl Acad. Sci. USA*, **102**, 5483–5488.

Wallace,E.W. *et al.* (2013) Estimating selection on synonymous codon usage from noisy experimental data. *Mol. Biol. Evol.*, **30**, 1438–1453.

Wang,M. *et al.* (2012) PaxDb, a database of protein abundance averages across all three domains of life. *Mol. Cell. Proteomics*, **11**, 492–500.

Warner,J.R. (1999) The economics of ribosome biosynthesis in yeast. *Trends Biochem. Sci.*, **24**, 437–440.

Wu,G. *et al.* (2008) Integrative analyses of posttranscriptional regulation in the yeast *Saccharomyces cerevisiae* using transcriptomic and proteomic data. *Curr. Microbiol.*, **57**, 18–22.

Wu,T.T. *et al.* (2009) Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, **25**, 714–721.

Zheng,Q. and Wang,X.-J. (2008) GOEAST: a web-based software toolkit for gene ontology enrichment analysis. *Nucleic Acids Res.*, **36(Suppl. 2)**, W358–W363.

Zheng-Bradley,X. *et al.* (2010) Large scale comparison of global gene expression patterns in human and mouse. *Genome Biol.*, **11**, R124.