Bioinformatics, 31(15), 2015, 2530–2536
doi: 10.1093/bioinformatics/btv182
Advance Access Publication Date: 29 March 2015
Original Paper

OXFORD

Systems biology

Outlier detection at the transcriptome-proteome interface

Yawwani Gunawardana[1], Shuhei Fujiwara[2], Akiko Takeda[2], Jeongmin Woo[3], Christopher Woelk[3] and Mahesan Niranjan[1,*]

[1]School of Electronics and Computer Science, University of Southampton, Southampton, UK, [2]Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan and [3]Faculty of Medicine, Southampton General Hospital, University of Southampton, Southampton, UK

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on November 30, 2014; revised on February 9, 2015; accepted on March 24, 2015

Abstract

Background: In high-throughput experimental biology, it is widely acknowledged that while expression levels measured at the levels of transcriptome and the corresponding proteome do not, in general, correlate well, messenger RNA levels are used as convenient proxies for protein levels. Our interest is in developing data-driven computational models that can bridge the gap between these two levels of measurement at which different mechanisms of regulation may act on different molecular species causing any observed lack of correlations. To this end, we build data-driven predictors of protein levels using mRNA levels and known proxies of translation efficiencies as covariates. Previous work showed that in such a setting, outliers with respect to the model are reliable candidates for post-translational regulation.

Results: Here, we present and compare two novel formulations of deriving a protein concentration predictor from which outliers may be extracted in a systematic manner. The first approach, outlier rejecting regression, allows explicit specification of a certain fraction of the data as outliers. In a regression setting, this is a non-convex optimization problem which we solve by deriving a difference of convex functions algorithm (DCA). With post-translationally regulated proteins, one expects their concentrations to be affected primarily by disruption of protein stability. Our second algorithm exploits this observation by minimizing an asymmetric loss using quantile regression and extracts outlier proteins whose measured concentrations are lower than what a genome-wide regression would predict. We validate the two approaches on a dataset of yeast transcriptome and proteome. Functional annotation check on detected outliers demonstrate that the methods are able to identify post-translationally regulated genes with high statistical confidence.

Contact: mn@ecs.soton.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The mapping between high-throughput measurements at the level of transcriptome and at the corresponding proteome is a complex one. Although a large body of computational biology literature using advanced machine learning algorithms to transcriptomic data exist, it is acknowledged that the underlying biological function of interest happens more at the protein level and mRNA concentrations are seen as proxies for the corresponding protein concentrations. Several authors have measured mRNA and protein concentrations in the same biological samples and have attempted to show

© The Author 2015. Published by Oxford University Press. All rights reserved. For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

correlations between these two (Gygi *et al.*, 1999; Futcher *et al.*, 1999; Marguerat *et al.*, 2012). Except under specific functional categories, correlation between the two is difficult to demonstrate. The reason for this is that different species of mRNA/proteins are regulated by different mechanisms at the post-transcriptional and post-translational levels.

The approach pursued in this work, starting from Tuller *et al.* (2007) and Gunawardana and Niranjan (2013), is to formulate a regression problem in which the response variable is the protein concentration and the covariates are the mRNA levels and other proxies for the stability and translation efficiency of the transcripts. Gunawardana and Niranjan (2013) used a linear regression model with a sparsity inducing regularizer (lasso) (Tibshirani, 1994) and showed that of about 37 features taken as inputs, a combination of five features including mRNA levels and translation efficiencies yield good prediction of protein levels. In fact, these best five features are mRNA abundance (Greenbaum *et al.*, 2003), tRNA adaption index (tAI), codon bias (Wall *et al.*, 2005), ribosome density and occupancy (Arava *et al.*, 2003). The outliers with respect to this linear regression were shown to carry significant over-representation of post-translationally regulated proteins, which is to be expected since the input covariates do not have any information about post-translational modifications (PTMs).

In this article, we introduce two novel computational methods that detect outliers in a regression setting and demonstrate their usefulness in the analysis of transcriptomic and proteomic datasets. Our focus is specific in that we attempt to detect from the data those proteins that are likely candidates for post-translational regulation. The relevant biological insight, introduced in Gunawardana and Niranjan (2013), is that if the concentration of a protein is regulated post-translationally, the measured abundance (P) of it is likely to be lower than what a global data-driven model trained on a genomic-wide scale might predict ($\hat{P}$). This is because the primary mechanism by which post-translational regulation might be implemented is the disruption of protein stability. However, other modifications which occur post-translationally, such as hydrophobicity, localization and enzymatic activities will not be detected by this approach.

The first novel model, we introduce in this article is outlier rejecting regression (ORR) model which formulates a regression problem that requires a user-specified fraction of the data to be returned as outliers with respect to the regression model. This is achieved via specifying a particular loss function, the clipped (truncated) loss which is shown to be equivalent to defining a certain fraction of the data as outliers (Xu *et al.*, 2006). We show how this objective function may be optimized via a difference of convex functions algorithm (DCA). For this formulation, we also present an alternative *ad-hoc* variant of optimization strategy (Methods). Our second method is the use of quantile regression (QR) with *omic* measurements, a technique which is effectively used in a range of areas including economics (Hendricks and Koenker, 1992; Koenker, 2005), medicine(Cole and Green, 1992; Heagerty and Pepe, 1999) and survival analysis (Koenker and Geling, 2001). QR enables the specification of an asymmetric loss function, where the user can define the outliers to be selected either with positive or negative losses. Thus, this approach is more suitable with our initial hypothesis because our main focus lies on the negative losses of a global predictor (measured abundance lower than the predicted—$P < \hat{P}$) to detect post-translationally regulated proteins. We believe that these are much neater ways of approaching the problem, than to simply implement a regression and hope the outliers to contain those post-translationally regulated genes.

2 Methods

In this article, we compare three types of outlier detection techniques at the transcriptome-proteome interface. Those are:

Model 0—Simple Linear Regression with the proteins lying further away from the regression line are considered as outliers. This method was used in Gunawardana and Niranjan (2013)'s study (previous work), where the regression was carried out with 37 input features with sparse inducing penalty (lasso) which selected five dominant features. In this study, all three models use these best five features as inputs: mRNA abundance, tAI, codon bias, ribosome density and occupancy,

Model 1—Novel ORR model, which explicitly formulates a user-defined proportion of the data as outliers. The model is estimated using two variants of constrained minimization algorithm (Section 2.1) which have similar convergence and outlier detection properties on the transcriptome-proteome problem,

Model 2—QR, allowing asymmetric loss functions to detect outliers only with negative losses (Section 2.2).

ORR (Model 1) and QR (Model 2) are the newly proposed methods and linear regression (Model 0) was only used to compare the new results with the previous work. Implementation details of the models are summarized later, with further derivations given in Supplementary Material (Section A) available with the online version of this article.

2.1 Outlier rejecting regression

Let $\{(\boldsymbol{x}_i, y_i)\}_{i=1,\ldots,m}$ where $\boldsymbol{x}_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$ be the set of m training samples. Our goal is to predict y as $f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b$. For any arbitrary convex loss function $\ell(\boldsymbol{x}, y; \boldsymbol{w}, b)$ such as hinge loss, squared loss and logistic loss, we define the *clipped loss function* using a variable η (e.g. Yang *et al.*, 2010; Wu and Liu, 2007) as below:

$$\begin{aligned}\ell_U(\boldsymbol{x}, y; \boldsymbol{w}, b) &:= \min\{\ell(\boldsymbol{x}, y; \boldsymbol{w}, b), U\},\\ &= \min_{0\le\eta\le 1} \eta\ell(\boldsymbol{x}, y; \boldsymbol{w}, b) + U(1-\eta),\end{aligned}$$

where the hyper parameter $U > 0$ to denotes the clipping threshold [Xu *et al.* (2006) for the second equality].

Here, we consider L2 regularized loss function in ridge regression (Equation 1) for the ORR model, which is similar to the squared loss $\ell(\boldsymbol{x}, y; \boldsymbol{w}, b) := \{y - (\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b)\}^2$, but with an additional regularization term to penalize data over-fitting and model complexity (Tibshirani, 1996; Andrew and Gao, 2007);

$$\min_{\boldsymbol{w},b} \lambda||\boldsymbol{w}||^2 + \{y - (\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b)\}^2, \tag{1}$$

where $\lambda > 0$.

Therefore, a clipped loss model for the ridge regression problem can be written as below:

$$\min_{\boldsymbol{w},b} \lambda||\boldsymbol{w}||^2 + \sum_i \ell_U(\boldsymbol{x}_i, y_i; \boldsymbol{w}, b), \tag{2}$$

where $U > 0$ and $\lambda > 0$ are hyper parameters.

It is troublesome to control U to define the outliers because this hyper parameter is highly depending on the error value distribution of the data samples. Therefore, we use a parameter which corresponds to the outlier ratio $\mu \in [0, 1)$ instead of U and consider the following model:

$$\begin{aligned}&\min_{\boldsymbol{w},b,\boldsymbol{\eta}} \lambda||\boldsymbol{w}||^2 + \frac{1}{(1-\mu)m}\sum_i \eta_i \ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}, b)\\ &\text{s.t.} \quad \textstyle\sum_i (1-\eta_i) \le \mu m, \;\; 0 \le \eta_i \le 1, \;\; \forall i,\end{aligned} \tag{3}$$

where $\mu \in [0, 1)$ and $\lambda \in (0, \infty)$ are hyper parameters.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

See Supplementary Material or Xu *et al.* (2006) for the relationship between Equations (2) and (3).

We believe the fraction of data to be expected as outliers is probably easier to specify using prior knowledge of the problem domain, than a threshold U on the regression loss.

Note that $\sum_i(1-\eta_i)=\mu m$ holds at the optimality. The sample $(\boldsymbol{x}_i, y_i)$ with $\eta_i^*=0$ can be regarded as an outlier for small $\mu>0$. However, this is a non-convex problem and finding a global solution for a non-convex problem is difficult.

2.1.1 Difference of convex functions algorithm

DCA (Pham Dinh and Le Thi, 1997; Collobert *et al.*, 2006) was employed to obtain plausible solutions for our non-convex problem. To solve the Equation (3), DCA updates $\boldsymbol{\eta}$ and $(\boldsymbol{w}, b)$ alternatively. We denote the solution at the kth iteration as $(\boldsymbol{w}^k, b^k, \boldsymbol{\eta}^k)$. $\boldsymbol{\eta}^k$ is computed using $(\boldsymbol{w}^k, b^k)$ as

$$\begin{aligned}&\boldsymbol{\eta}^k \in \operatorname{argmax}_{\boldsymbol{\eta}} \textstyle\sum_i(1-\eta_i)\ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}^k, b^k)\\ &\text{s.t.}\textstyle\sum_i(1-\eta_i)=\mu m,\ \ 0\le\eta_i\le 1.\end{aligned} \tag{4}$$

We can find $\boldsymbol{\eta}^{\mathrm{k}}$ by sorting the losses and assigning 0 to η_i^k with large loss. Hence, the computational cost is very small. $(\boldsymbol{w}^{k+1}, \boldsymbol{b}^{k+1})$ is also computed using $\boldsymbol{\eta}^{\mathrm{k}}$ as a solution of the following convex problem:

$$\min_{w,b}\ \lambda||\boldsymbol{w}||^2+\frac{1}{(1-\mu)m}\ \left[\sum_i \ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}, b)-\mu m(\langle \boldsymbol{g}_w, \boldsymbol{w}\rangle + g_b b)\right] \tag{5}$$

where

$$\boldsymbol{g}_w=\frac{1}{\mu m}\sum_i(1-\eta_i^k)\nabla_w\ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}^k, b^k), \tag{6}$$

$$g_b=\frac{1}{\mu m}\sum_i(1-\eta_i^k)\nabla_b\ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}^k, b^k). \tag{7}$$

For a non-smooth convex loss function $\ell(.)$, let $\nabla_w\ell(.)$ and $\nabla_b\ell(.)$ denote subgradients of $\ell(.)$ with respect to $\boldsymbol{w}$ and b. The sequence $(\boldsymbol{w}^k, b^k)$ generated by Algorithm 1 has the following good convergence properties: the objective value of Equation (3) is decreasing and every limit point of the sequence is a critical point satisfying a necessary condition for local minima of (3).

Algorithm 1 shows the DCA approach in a pseudo-code format. This was implemented using CVX package in MATLAB environment.

2.1.2 Alternative heuristic implementation of DCA

We also used an alternative heuristic implementation of DCA in ORR model (again using CVX package in MATLAB) as shown in Algorithm 2. λ and μ values were set similar to the Algorithm 1 ($\lambda=0.01$ and $\mu=0.975$). We observed that this implementation also selects the same set of proteins as similar to Algorithm 1. Thus, these two algorithms produce identical results.

Algorithm 2 is similar to Algorithm 1, but easier and more intuitive. The difference between Algorithm 1 and 2 is the step where a subproblem is solved. Algorithm 2 solves Equation (3) with respect to $(\boldsymbol{w}, b)$ by fixing η. Therefore, we solve the following problem:

$$\min_{w,b}\ \lambda||\boldsymbol{w}||^2+\frac{1}{(1-\mu)m}\sum_i \eta_i^k\ell(\boldsymbol{x}_i, y_i; \boldsymbol{w}, b). \tag{8}$$

MATLAB scripts for Algorithm 1 (function ORR1) and 2 (function ORR2) are shown in Supplementary Material (Section A).

2.2 Quantile regression

Here, we consider different weights for the negative and positive losses given by $y-(\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b)$, for all i. $\tau\in(0,1)$ determines the quantile of interest and $\tau=0.5$ represents the symmetric error with conditional median (Koenker, 2005).

Quantile loss is given as $\rho_\tau(y-(\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b))$ where

$$\rho_\tau(z)=\begin{cases}\tau\cdot(z) & \text{if } (z)\ge 0,\\ -(1-\tau)\cdot(z) & \text{otherwise.}\end{cases} \tag{9}$$

We can obtain the outliers of our interest [i.e. in our case $\{y-(\langle \boldsymbol{w}, \boldsymbol{x}\rangle + b)\}<0$] using Equation (10) by setting the τ to the required quantile.

$$\min_{w,b}\ \sum_i\{\rho_\tau(y_i-(\langle \boldsymbol{w}, \boldsymbol{x_i}\rangle + b))\} \tag{10}$$

Equation (10) can be solved as a linear program as shown in Equation (11). We used *linprog* function in MATLAB environment to implement Equation (11).

$$\begin{aligned}\min_{w,b}\ \ &\sum_i\{\tau u_i+(1-\tau)v_i\}\\ \text{s.t.}\ \ &y_i-(\langle \boldsymbol{w}, \boldsymbol{x_i}\rangle + b)=u_i-v_i, \forall i\\ &u_i\ge 0,\ \ v_i\ge 0, \forall i\end{aligned} \tag{11}$$

MATLAB function call for this optimization is given in Supplementary Material (Section A).

2.3 Post-translational regulation annotation check

Similar to the previous work (Gunawardana and Niranjan, 2013), functional annotation check was carried out at two levels (i.e. coarse and finer levels). At the coarse level, PTMs keywords obtained by UniProt database (Magrane and Consortium, 2011) were consider as the only requirement to indicate post-translation regulation and at the finer level, PTMs coupled with the motif information (i.e. Phosphorylation + PEST motifs, Acetylation + N-termini segments and Ubiquitination + D or KEN Box motifs) which are directly affecting the protein stability were considered as more powerful indicators to post-translation regulation. EMBOSS explorer epestfind (Rice *et al.*, 2000), NetAcet 1.0 (Kiemer *et al.*, 2005) and GPS-ARM 1.0 toolkit (Liu *et al.*, 2012) databases were used to obtain motif information for PEST, N-termini segment and D/Ken box motifs, respectively.

Investigating the statistical significance of the number of post-translationally regulated proteins detected in the outlier region with respect to all other proteins was measured by obtaining 1000 random samples with sample size 50 (2.5% from the total dataset), where the sampling process can be considered as computationally exhausting.

We also carried out a gene enrichment analysis on the outlier sets to uncover useful biological insights. Gene ontology (GO) analysis of outliers was performed using BiNGO 2.44, a plugin for Cytoscape (Maere *et al.*, 2005) and PANTHER web tool by Thomas *et al.* (2003) was used for pathway analysis. We also used WebGestalt tool which draws from multiple large databases to explore more gene enrichment properties (Zhang *et al.*, 2005). Protein-protein interactions at the physical layer was carried out using BioGRID database (Stark *et al.*, 2006) and GeneMANIA web tool was employed to discover more biological network relationships among the outlier proteins (Warde-Farley *et al.*, 2010).

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Algorithm 1 DCA for Outlier Rejecting Regression

Require: Initial $(\boldsymbol{w}^0, b^0)$; hyper-parameters $\mu \in [0,1)$ and $\lambda \in (0,\infty)$.
$k \leftarrow 0$.
repeat
Obtain $\boldsymbol{\eta}^{\mathrm{k}}$ of (4) by sorting $\ell(\boldsymbol{x}_{\mathrm{i}}, y_i; \mathrm{w}^k, b^k), \forall i$.
Computer $(\boldsymbol{g}_{\mathrm{w}}, g_b)$ using $\boldsymbol{\eta}^{\mathrm{k}}$ as (6) and (7).
$(\boldsymbol{w}^{\mathrm{k}+1}, b^{k+1}) \leftarrow$ a solution of subproblem (5).
$k \leftarrow k+1$.
until convergence.

Algorithm 2 Alternative Heuristic Implementation of Outlier Rejecting Regression

Require: Initial $(\boldsymbol{w}^0, b^0)$; hyper-parameters $\mu \in [0,1)$ and $\lambda \in (0,\infty)$.
$k \leftarrow 0$.
repeat
Obtain $\boldsymbol{\eta}^{\mathrm{k}}$ of (4) by sorting $\ell(\boldsymbol{x}_{\mathrm{i}}, y_i; \boldsymbol{w}^k, b^k), \forall i$.
$(\boldsymbol{w}^{\mathrm{k}+1}, b^{k+1}) \leftarrow$ a solution of subproblem (8).
$k \leftarrow k+1$.
until convergence.

3 Results

3.1 Regression

With the five selected input features [covariates chosen in the previous work, Gunawardana and Niranjan (2013)], all three regression models achieved a good level of predicting out of sample protein levels ($R^2 = 0.86$ for simple linear regression, $R^2 = 0.86$ for ORR and $R^2 = 0.85$ for QR). Figure 1 shows the three prediction results as scatter plots [predicted ($\hat{P}$) against measured (P) concentration] with the detected outlier points are shown as circles. We also compared the outputs of ORR and QR models to confirm that both produce correlated results. Figure 2 illustrates that these two new models produce highly correlated results with $R^2 = 0.97$. Additionally, Supplementary Figure S2 shows the correlation of ORR and QR model outputs with respect to the simple linear regression model and we observed that all three models produced highly correlated outputs.

Afterwards, these three sets of outliers were tested at two levels (coarse and finer) of functional annotations to obtain evidence for post-translational regulation. Table 1 shows that all three regression models have high level of confidence (P-values < 0.05) to support our hypothesis at both levels. Note that QR model gives the highest confidence level to detect post-translationally regulated proteins as outliers.

3.2 Validating ORR and QR models

Two synthetic datasets were employed to validate the outlier detection by ORR and QR models. Hawkins *et al.* (1984)'s artificial dataset was used to detect outliers using ORR model which contained two groups of outliers with positive and negative losses, respectively. Boston Housing Data from UCI Machine Learning Repository (Bache and Lichman, 2013) was employed to detect one side (positive loss) outliers using QR model. Supplementary Figure S3(A) shows that ORR was able to detect all 14 outliers (both Group 1 and 2) while separating them into two clusters. Similarly QR model [Supplementary Figure S3(B)] was able to detect top 20 outliers only with positive losses. See Supplementary Section C for further details.

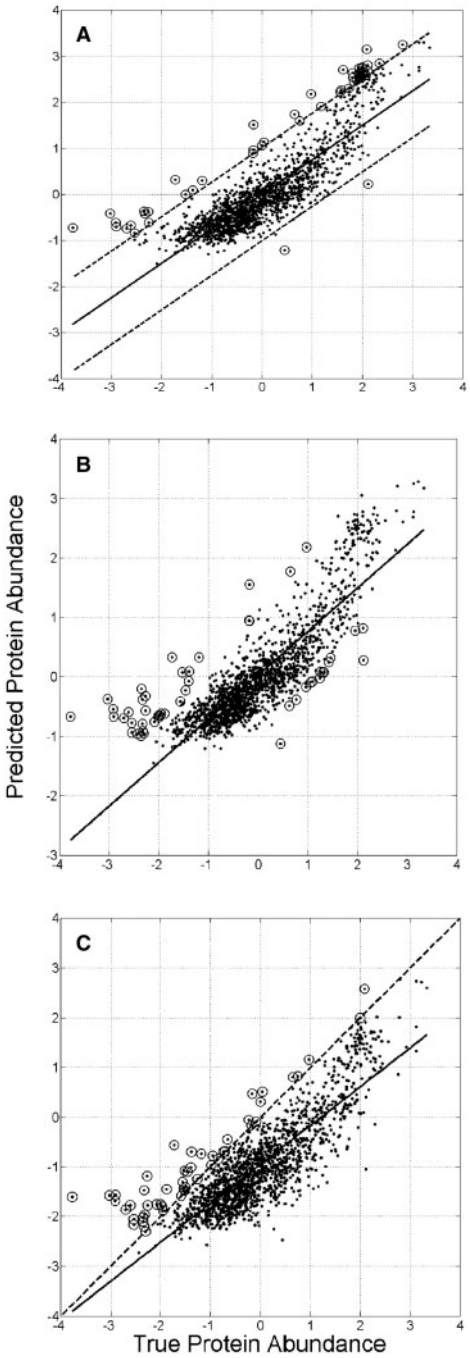

Fig. 1. Outlier detection by three regression models. Outliers detect by three regression models on scatter plots of predicted and true protein abundance (in log scale) and the detected outliers are shown as circles. (**A**) Simple Linear Regression model used by Gunawardana and Niranjan (2013), dash lines showing threshold set so that 2.5% of the data (50 proteins) are selected as outliers. (**B**) ORR model introduced in this article with 50 data points forced to be outliers. This model distributes outliers on either side of a best fitting regression line. (**C**) QR model which allows an asymmetric loss function enables the detection of outliers with true abundance lower than a global predicted—the biological insight we wish to impose on the models

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

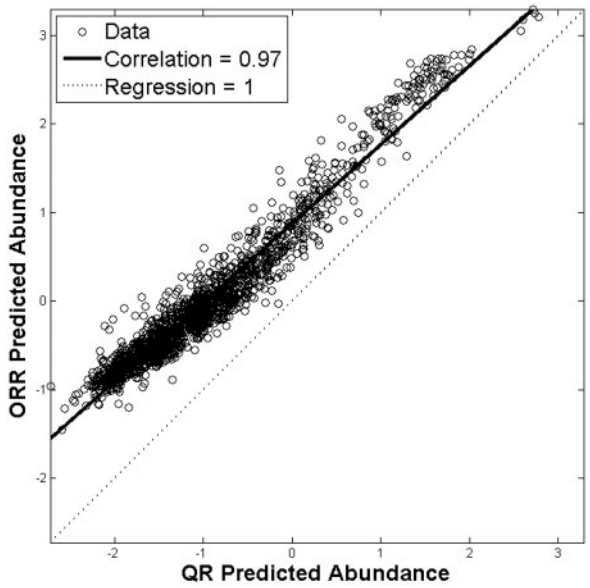

Fig. 2. Comparison of prediction from ORR and QR regression models (Methods). Strong global agreement ($R^2 = 0.97$) in model fit is obtained, but the models detect different sets of data as outliers (Fig. 3, Venn diagram for overlap) due to the nature of loss function imposed

3.3 ORR convergence speed

We also compared the convergence speed of Algorithm 1 and 2 of ORR model using four different datasets (Supplementary Section D). Supplementary Figure S4 shows, in all cases, Algorithm 2 converges faster than Algorithm 1. Note that the convergence speed also depends on the size of the dataset. Though, the speed is not a major factor with respect to our problem (transcriptome-proteome data), Algorithm 2 may well be a better implementation for large regression problems.

3.4 Biological interpretation of outliers

Figure 3 shows the distribution of outlier genes of the three regression models in a Venn diagram. Ninety-two and 17 genes were found as the union and the intersection of the Venn diagram, respectively. Here, we discuss biological aspects of these outliers using GO enrichment and pathway analysis.

Gene enrichment analysis carried out on the union of 92 genes found by all three models showed several GO terms and pathways related to translational and post-translational regulation. These are shown in Supplementary Table S1 and Figure S5. Ribosomal GO terms were dominant in biological process category and WebGestalt tool also found several keywords related to ribosomal proteins (Supplementary Table S2). It is known that, ribosomal proteins undergo several post-translational regulation such as N-terminal acetylation, removal of methionine, phosphorylation and N-terminal methylation (Carroll *et al.*, 2008). However, previous work (Gunawardana and Niranjan, 2013) discusses that ribosomal proteins did not unduly influence the over-representation of post-translational regulation of the outliers in a global regression model.

We subjected the consensus (intersections) and the union genes for pathway analysis. Interestingly in all cases (considering two models at a time and all three), the *p*53 pathway and *p*53 pathway feedback loop contained significant over-representation of outlier genes. Shin *et al.* (2013) showed that protein degradation mechanism of post-translational regulation enables robust *p*53 regulation by stabilizing the *p*53 levels with less energy. In fact, during DNA damage conditions, Mdm2 feedback loop down-regulates *p*53 levels by the aid of the protein degradation process of post-translational regulation (Šmardová *et al.*, 2005; Shin *et al.*, 2013). Thus, the over-representation of outlier genes in the *p*53 related pathways re-confirmed our hypothesis by providing evidence that the outliers are more likely to be post-translationally regulated.

Table 1. Function annotation check for the 50 outliers taken from each regression model

Regression model		Coarse level		Finer level
	No of genes	Confidence level	No of genes	Confidence level
Simple linear regression (Model 0)	44	$P < 2.00 \times 10^{-02}$	37	$P < 2.11 \times 10^{-10}$
ORR (Model 1)	40	$P < 4.80 \times 10^{-02}$	35	$P < 8.31 \times 10^{-09}$
QR (Model 2)	45	$P < 9.90 \times 10^{-04}$	38	$P < 2.94 \times 10^{-11}$

One thousand random trials were used obtain the *P*-values

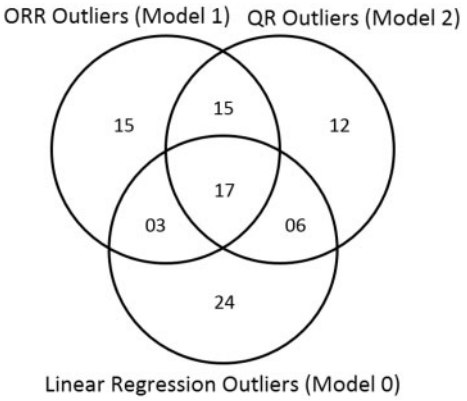

Fig. 3. Venn diagram showing the number of outlier proteins that overlap between the three prediction models

Further, protein interaction network was constructed using BioGRID interactome database to identify physical interactions of protein product from outlier genes (union 92 genes) defined by our three models (Supplementary Figure S6). Large cluster of this network includes ribosomal subunits such as RPS16B, RPS17B, RPS13, RPS14A, RRP2B, RPL17A, RPL17B and RPS9B, which were mostly selected as outliers from simple linear regression model (Model 0). As mentioned earlier, ribosomal proteins have efficient translational activities (Warner, 1999), followed by several post-translational regulation (Carroll *et al.*, 2008). In addition to the main cluster of ribosomal subunit components, another protein cluster with hub PHO88, the protein which are known to be involved with phosphate ion transport and protein maturation (Čopič *et al.*, 2009), was identified and this is a process which involves in phosphorylation PTM to adapt phosphate group to the protein (Burnett and Kennedy, 1954). Supplementary Figure S7 shows co-expression, genetic, predicted and physical interaction networks obtained by GeneMANIA web tool. Co-expression network gave the highest coverage of the related proteins. However, it did not show any statistical significance of our union outlier set with respect to random samples (Supplementary Section F). We also observed that, with all four networks, ribosomal proteins tend to cluster together (purple colour nodes in Supplementary Figure S7) and the physical interactions are similar to BioGRID output.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Additionally, we also carried out a finer level functional annotation check, GO and pathway analysis on the common genes (intersection of two models and all three) to discuss more biological insights of these outlier genes (Supplementary Section E). We observed an over-representation of PTM functional annotations and *p*53 related pathways with all the combinations of outlier sets.

4 Discussion

Tests against functional annotation of yeast gene products show that all three methods can detect outlier proteins that are likely candidates for post-translational regulation with high statistical confidence. However, main biological insight that we start from is that post-translational regulation should primarily act by disrupting the stability of proteins whereby the measured concentrations should be lower than what might be predicted from a genome-wide regression with mRNA level properties of individual species as input variables. Two of the models we considered (simple linear regression and ORR) do not model this asymmetric explicitly and hence find outliers on either side (i.e. $P < \hat{P}$ and $P > \hat{P}$). In linear regression, the vast majority of outliers found had $P < \hat{P}$, in line with our insight, and only two proteins were found as outliers on the $P > \hat{P}$ side. This has to be regulated as a prediction result obtained by fitting a data-driven model through noisy data. When we force the model to label a fraction of the data as outliers in ORR model, the minimization of the loss function picks up outliers on either side (i.e. $P < \hat{P}$ and $P > \hat{P}$). However, majority of the outliers and the post-translatonally regulated proteins were detected from the upper region ($P < \hat{P}$) of the regression. When we compared the outliers detected on either side of ORR model, we found 7 of the 15 outliers detected at the lower region ($P > \hat{P}$) did not have post-translational regulation, and correspondingly the confidence levels with which ORR model identify post-translationally regulated proteins was lower. The QR model, which allows as to explicitly impose our asymmetric loss function turns out to be the best match to exploit the biological insight we purse here, and the proteins picked up by this model yield the highest statistical confidence in annotation checks. Pathway analysis further supports the hypothesis due to the over-representation observation of *p*53 related pathways.

5 Conclusion

In this work, we presented two novel approaches to detect post-translationally regulated proteins as outliers in a regression problem. The novelty of the computational formulations we explore in this work lies in (i) the explicit formulation that a certain fraction of data should be detected as outliers, and (ii) the error to be minimized being one sided because the measured concentration of post-translationally regulated proteins are expected to be lower than what a global regression would predict. Both these are ways of capturing our prior knowledge of the problem domain in the computational formulation. The methods we propose are shown to have the power to identify proteins whose stability is disturbed by post-translationally acting processes to a statistical significance, thus these data-driven techniques help to uncover functional aspects of molecular biology.

Conflict of Interest: none declared.

References

Andrew,G. and Gao,J. (2007) Scalable training of L1-regularized log-linear models. In: Proceedings of the 24th International Conference on Machine Learning, ACM, pp. 33–40.

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

Bache,K. and Lichman,M. (2013) UCI Machine Learning Repository **901**, http://archive.ics.uci.edu/ml.

Burnett,G. and Kennedy,E.P. (1954) The enzymatic phosphorylation of proteins. *J. Biol. Chem.*, **211**, 969–980.

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.

Cole,T.J. and Green,P.J. (1992) Smoothing reference centile curves: the LMS method and penalized likelihood. *Stat. Med.*, **11**, 1305–1319.

Collobert,R. *et al.* (2006) Trading convexity for scalability. In: International Conference on Machine Learning, ACM, pp. 129–136.

Čopič,A. *et al.* (2009) Genomewide analysis reveals novel pathways affecting endoplasmic reticulum homeostasis, protein modification and quality control. *Genetics*, **182**, 757–769.

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

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

Gunawardana,Y. and Niranjan,M. (2013) Bridging the gap between transcriptome and proteome measurements identifies post-translationally regulated genes. *Bioinformatics*, **29**, 3060–3066.

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

Hawkins,D.M. *et al.* (1984) Location of several outliers in multiple-regression data using elemental sets. *Technometrics*, **26**, 197–208.

Heagerty,P.J. and Pepe,M.S. (1999) Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in US children. *J. R. Stat. Soc. Ser. C (Appl. Stat.)*, **48**, 533–551.

Hendricks,W. and Koenker,R. (1992) Hierarchical spline models for conditional quantiles and the demand for electricity. *J. Am. Stat. Assoc.*, **87**, 58–68.

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

Koenker,R. (2005) *Quantile Regression*. **Vol. 38**. Cambridge University Press, Cambridge.

Koenker,R. and Geling,O. (2001) Reappraising medfly longevity: a quantile regression survival analysis. *J. Am. Stat. Assoc.*, **96**, 458–468.

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.

Maere,S. *et al.* (2005) BiNGO: a cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, **21**, 3448–3449.

Magrane,M. and Consortium,U. (2011) Uniprot knowledgebase: a hub of integrated protein data. *Database*, doi: 10.1093/database/bar009.

Marguerat,S. *et al.* (2012) Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. *Cell*, **151**, 671–683.

Pham Dinh,T. and Le Thi,H.A. (1997) Convex analysis approach to D.C. programming: theory, algorithms and applications. *Acta Math. Vietnamica*, **22**, 289–355.

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

Shin,Y. *et al.* (2013) Post-translational regulation enables robust p53 regulation. *BMC Syst. Biol.*, **7**, 83–94.

Šmardová,J. *et al.* (2005) Functional analysis of p53 tumor suppressor in yeast. *Differentiation*, **73**, 261–277.

Stark,C. *et al.* (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**(Suppl. 1), D535–D539.

Thomas,P.D. *et al.* (2003) PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.*, **13**, 2129–2141.

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

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Methodological)*, **58**, 267–288.

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

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

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

Warde-Farley,D. *et al.* (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.*, **38**(Suppl. 2), W214–W220.

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

Wu,Y. and Liu,Y. (2007) Robust truncated hinge loss support vector machines. *J. Am. Stat. Assoc.*, **102**, 974–983.

Xu,L. *et al.* (2006) Robust support vector machine training via convex outlier ablation. In: *American Association for Artificial Intelligence (AAAI)*, (www.aaai.org) pp. 536–542.

Yang,M. *et al.* (2010) Relaxed clipping: a global training method for robust regression and classification. In: *Neural Information Processing Systems, Curran Associates, Inc.* pp. 2532–2540.

Zhang,B. *et al.* (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res.*, **33**(Suppl 2), W741–W748.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016