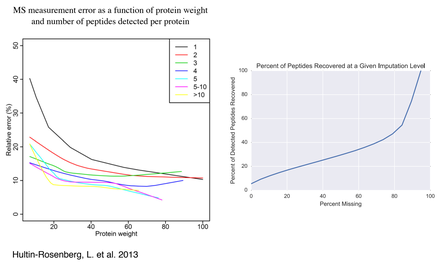
Applications of Machine Learning in the Processing and Analysis of Large Circadian Proteomics Time-Series Datasets

Alexander Crowell[1], Jennifer Hurley[2], Erika Bredeweg[3], Erika Zink[3], Samuel Purvine[3], Scott Baker[3], Jennifer Loros[4], and Jay Dunlap[1]

[1]Dept. of Genetics, Geisel School of Medicine, Dartmouth, Hanover, NH, USA [2]Dept. of Biological Science Rensselaer Polytechnic Institute, Troy, NY, USA
[3]Pacific Northwest National Laboratories, Richland, WA, USA [4]Dept. of Biochemistry, Geisel School of Medicine, Dartmouth, Hanover, NH, USA



Hultin-Rosenberg, L. et al. 2013

Background

High throughput techniques are becoming a staple of circadian analyses. Unlike RNAseq in which all transcripts are potentially identified, proteomics analysis by tandem mass spectrometry (MS) produces measurements in which the failure to observe a peptide is partly a factor of its abundance and partly effectively random. Additionally, the large number of samples required for circadian analyses necessarily introduces between run variation resulting in complex and arbitrary biases. We have developed methods to circumvent these inherent obstacles. A test case involves analysis of circadian proteomics data derived from time courses of *Neurospora crassa*. This dataset encompasses 48 hours with a two hour resolution, three replicates and two genotypes for a total of 144 samples. In MS only a small number of peptides are detected in all samples yet many are detected in almost all. Since most methods for analysis of rhythmicity require complete datasets this random missingness can greatly reduce the usefulness of data; however, the number of proteins detected and the reliability of abundance estimates can be greatly improved by imputing the values of partially missing peptides. We accomplish this using the K nearest neighbors method, which uses the mean of the K nearest data points. In order to remove the bias trends introduced by MS run variation we employ SVD modelling of residuals, which takes a known pattern of replication within the dataset, in this case biological and circadian time replicates, and models variation not shared between those replicates as trends which can then be removed from the data. In the case of the dataset described here, this method has been quite successful. The K Nearest Neighbors imputation has allowed us to recover 29,032 peptides, approximately doubling the number of peptides recorded from the original value of 30,681. The SVD normalization removed major bias trends associated with date of isobaric labeling and MS run order which otherwise dominated the circadian signal. In the wild type set ~ 13% of detected and ~6% of total proteins were observed to be circadian.

Algorithm 1 K Nearest Neighbors Imputation

procedure KNN($\mathcal{D}$,**K**)

$\mathcal{D}^{n\times m} \leftarrow$ Our dataset
$n \leftarrow$ number of observed peptides
$m \leftarrow$ number of MS experiments (timepoints × replicates)
$d \leftarrow$ a datum (all observations on a peptide)

for $d \in \mathcal{D}$
find **K** nearest neighbors of d by Euclidean distance

construct the matrix $\mathcal{N}^{\mathbf{K}\times m}$ with rows being nearest neighbors of d

impute missing values in d as mean of corresponding columns of $\mathcal{N}$
end

return complete $\mathcal{D}$

end procedure

Algorithm 2 SVD Bias Modelling

procedure SVD_Bias($\mathcal{D}, \alpha$)

$\mathcal{D}^{n\times m} \leftarrow$ Our dataset
$n \leftarrow$ number of observed peptides
$m \leftarrow$ number of MS experiments (timepoints × replicates)
$\alpha \leftarrow$ bias trend significance threshold
$\mathbf{S} \leftarrow$ Sample timepoints
$c_t \leftarrow$ correlation threshold

calculate the correlation to the primary variable of interest for each peptide c

form a reduced data matrix $\mathcal{D}_r$ of peptides for which $c < c_t$

fit the Lowess model $\mathcal{D}_r = \beta\mathbf{S}^T + \mathbf{E}$

Calculate the residual matrix as $\hat{\mathbf{E}} = \mathcal{D}_r - \beta\mathbf{S}^T$

Calculate the singular value decomposition of the residual matrix $\hat{\mathbf{E}} = \mathbf{UDV}^T$

With d_i as the i^th singular value, for right singular value k = 1, . . . , n calculate the observed test statistic as:

$$T_k = \frac{d_k^2}{\sum_{i=1}^{n} d_i^2}$$

Permute each row of the matrix $\hat{\mathbf{E}}$ independently to form a matrix $\hat{\mathbf{E}}^*$

Calculate the singular value decomposition of the matrix $\hat{\mathbf{E}}^* = \mathbf{U_0 D_0 V_0^T}$

For right singular value k calculate the null statistic:

$$T_k^0 = \frac{d_{0k}^2}{\sum_{i=1}^{n} d_{0i}^2}$$

Repeat calculation of the null statistic B times

Calculate the p value for right singular vector k as:

$$p_k = \frac{\#\{T_k^{0b} > T_k, b=1,\ldots,B\}}{B}$$

Estimate number of significant trends sb as:

$$sb = \sum_{k=1}^{n} \mathcal{I}(p_k \leq \alpha) \textbf{ While } p_k \leq \alpha$$

For each significant trend $\mathbf{v}_k$, regress $\mathbf{v}_k$ on each row of $\mathcal{D}_r$ calculating a p-value for their association

Estimate the number of truly associated features as $\hat{m}_1 = [(1-\hat{\pi}_0)\times m]$ and form a subset of features with the $\hat{m}_1$ smallest p-values

Calculate the right singular vectors of the reduced subsetted matrix as $\mathbf{v}_j^r$ for $j = 1, \ldots, n$

Estimate the surrogate variable j^* as the eigengene of the reduced subset matrix most correlated with the corresponding residual eigengene
$j^* = argmax_{1<j<n}\mathbf{cor}(\mathbf{v}_k, \mathbf{v}_j^r)$

Set $\hat{\mathbf{G}}_k = \mathbf{v}_{j^*}^r$.

return $\mathcal{D} - \mathcal{D} \times \hat{\mathbf{G}}^{-1} \times \hat{\mathbf{G}}$

end procedure

Imputation of Missing Values

A large fraction of peptides present are not detected in a given MS experiment, however the high number of time points and replicates common to circadian time courses mean that a significant portion of peptides are detected in most but not all experiments in a time course. Since most circadian classification software (JTK,eJTK) cannot handle missing values, it is important to impute these missing values so as not to waste valuable information. Since the failure to detect a given peptide is some unknown function of both its abundance and effectively random factors however, it is not meaningful to simply impute these missing values as 0. A large amount of study has gone into imputation algorithms across a variety of fields and while there are many options, we will describe only the one used here, K nearest neighbors (KNN). The KNN method has the advantage of being non-parametric, fast and highly reliable at the imputation levels and dataset sizes described here, making it the most popular choice for imputation in modern machine learning. The KNN algorithm imputes missing data by finding the K nearest data points with complete data for a given data point with missing data in the non-missing dimensions and imputes the missing value as the average of the corresponding values from the k nearest neighbors. By using this method with a relatively conservative imputation threshold of 30% we were able to double the size of our dataset, significantly improving both the number of proteins detected as well as the quality of our abundance estimates for proteins for which additional peptides were detected.

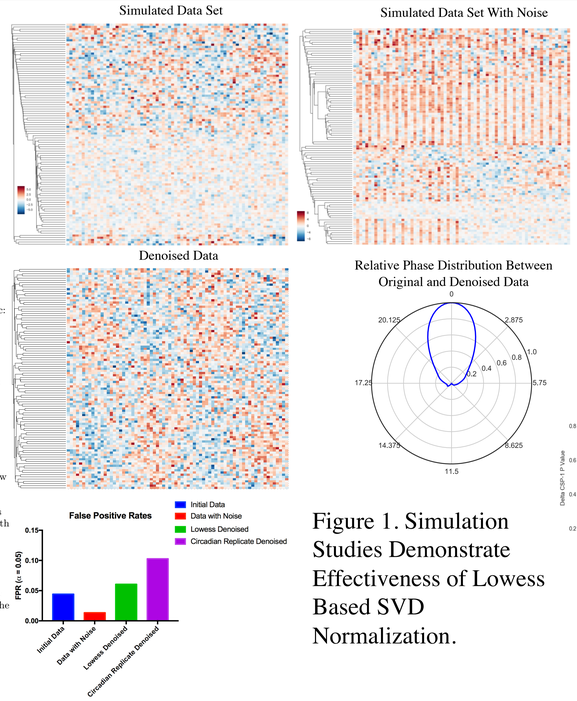
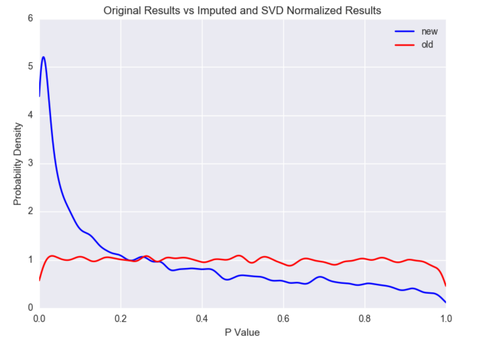


Figure 1. Simulation Studies Demonstrate Effectiveness of Lowess Based SVD Normalization.

Modelling of Bias Trends

MS experiments, are subject to more complex and larger systematic biases than RNAseq. Because of the large number of time points and replicates required for circadian time courses, as well as the relatively subtle trends we are attempting to detect, these biases can easily swamp any meaningful information from the sort of experiment described here. A lowess model can easily isolate the component of variation in a given peptide which is consistent between biological replicates and nearby timepoints separating it from the model residuals. Completely eliminating the model residuals would drastically underestimate our uncertainty however, so we must instead borrow from the discipline of unsupervised machine learning in order to gain a more nuanced picture of the residual matrix. The technique employed here to accomplish that goal is Singular Value Decomposition (SVD). SVD has been used to model bias trends in both micro-array and MS data. In general, SVD can be employed to break down the residual matrix into singular vectors (analogous to principle components) which each contribute independent bias trends. By repeatedly permuting the residual matrix and performing SVD, we can then estimate the null distribution of variance explained by each singular vector and thereby assess the 'significance' of individual bias trends, removing only those which explain more variance than we would expect by chance. The algorithm employed here is based on earlier work (Leek & Storey 07) but notable both for being a modern implementation in python and for implementing a more advanced reduced subset method originally proposed by Leek in tandem with Lowess based calculation of residuals.



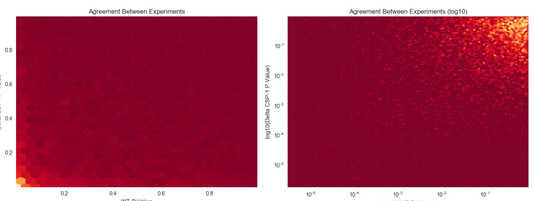


Figure 2. Improved Recovery of Circadian Proteins and Correspondence Between Experiments.

Results

The combination of KNN based imputation and SVD based error modeling improved our results markedly. Before the application of these techniques the p value distribution from eJTK was indistinguishable from what would be expected with a random input and afterward it showed clear enrichment for circadian proteins. Simulation Studies demonstrate the superior results from Lowess based residual calculations with low FPR, excellent denoising and effective recovery of phase. When comparing between the two proteomics datasets described here (WT and delta CSP-1) as well as an earlier RNAseq time course (Hurley et al. 2014), it is also clear that the predominant class is genes/proteins which are circadian in both pairs of datasets being compared (point of highest density lower left corner). This is the pattern we would expect if we were enriching for truly circadian genes and provides strong support for our methods. We would expect true negatives to be drawn from a uniform distribution and therefore genes/proteins which are circadian in neither of the conditions being compared would be expected to produce a low density background across the range of possible values, just as is observed. Additionally, the log-log comparison of p values shows the two values to be correlated which is supported by tests of correlation (Spearman [rho=0.18,p=7e-35], Chi Sq [p=1e-25]). There is also strong agreement between the phases of genes circadian in both proteomics experiments, which one would not expect if proteins were being artificially enriched as circadian. Both of these patterns fit well with biological expectations further supporting our methods

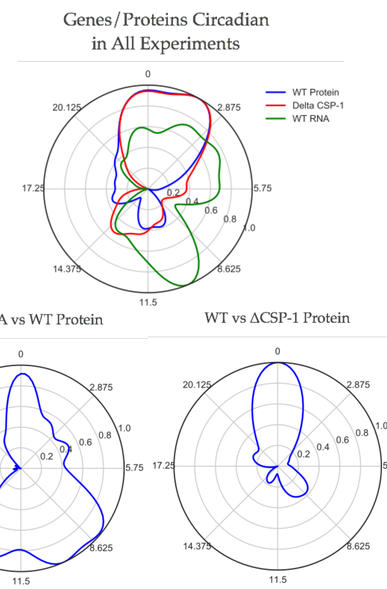


Figure 3. Novel RNA and Protein Phase Relationships Revealed Post Normalization

Conclusion

The application of imputation and bias modelling to this dataset was very successful and demonstrates the potential these techniques hold for the analysis of large scale circadian proteomics time courses. A point worth noting is that randomization of samples during MS as well as a high number of time points and replicates are critical to this type of analysis and while such large experiments require the investment of additional resources, their importance to this type of study cannot be overemphasized. Although there is still room for improvement in the design of circadian proteomics experiments both in vitro and in silico, this work makes clear the value of more sophisticated analysis techniques for such datasets. . The significance of these improvements can be seen in the novel phase relationships between circadian RNA and protein revealed by this work.

While past work has shown 30% imputation to be stable and effective, this threshold still results in discarding more than half of detected peptides. Therefore, one area for further study would be the analysis of more computationally intensive imputation methods which could potentially impute a larger fraction of missing data.

ACKNOWLEDGEMENTS

Funding was provided by NIH (NIGMS) grants to J. Loros (R01 GM083336) and J. Dunlap (R01 GM096574) and also by the DOE (PNNL) to J. Dunlap, J. Hurley and J. Loros (47818). Thanks to the FGSC for strains and support.