Quantifying Uncertainty in Mass Spectrometry Based Spatial Proteomics

Oliver M. Crook, Laurent Gatto, Paul D.W.Kirk, Kathryn Lilley

Cambridge Centre for Proteomics, Computational Proteomics Unit, MRC Biostatistics Unit, University of Cambridge

Introduction

- The sub-cellular localisation of proteins is crucial to execute their intended function, and aberrant localisations are a hallmark of many diseases, including cancer and obesity.
- State-of-the art experimental procedures exist, that rely on separation of cellular content and high accuracy mass spectrometry, to determine protein localisations.
- In the *hyper*LOPIT protocol, organelles and macro-molecular complexes are characterised by density-specific profiles along a gradient.
- Quantitative protein profiles that match the organelle profiles along the gradient are produced using high throughput mass spectrometry (MS).
- Organelle profiles can be modelled using non-parametric Bayesian techniques such a Gaussian Process regression.

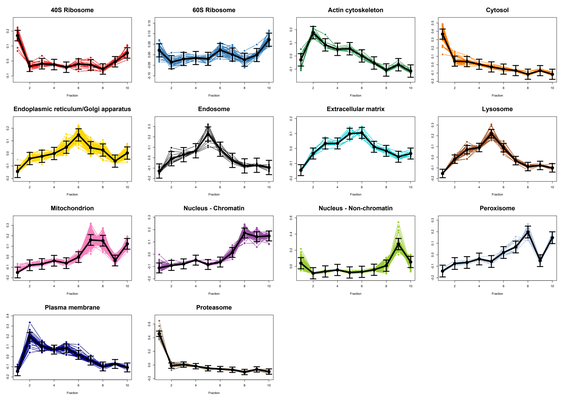

Figure 1:Proteins with known localisation to organelles and macro-molecular complexes in a mouse pluripotent embryonic stem cell dataset are modelled using Gaussian processes. The Gaussian process regression model captures the non-linearity of each unique profile.

- Previously, supervised machine learning algorithms have been employed to create classifiers that make protein-organelle assigments.
- However, proteins can be distributed amongst multiple localisations and trans-locate upon perturbation by external stimuli, leading uncertainty which we wish to quantify.
- We propose a semi-supervised non-parametric Bayesian framework to create a generative model to classify proteins to sub-cellular niches and quantify the uncertainty in our assignments.

Methods

- We model our data as a finite mixture of Gaussian process regression models.
- In the presence of outlier we introduce an additional component to the mixture model, which takes the form of the Student's t-distribution because heavy tailed distributions are good at capturing dispersed proteins.
- Equation 1 captures the full complement of proteins, where π_k are our mixture weights, F denotes the density of the Gaussian Process and G the density of the Student's t-distribution and ϕ_i denotes an indicator to the outlier component.

$$p(\mathbf{x}_i|\boldsymbol{\pi},\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k F(\mathbf{x}_i|\boldsymbol{\theta}_k)^{\phi_i} G(x_i|\Phi)^{1-\phi_i}. \quad (1)$$

- We take a Fully Bayesian approach to inference, placing standard normal hyper-priors on the log-hyperparameters of the Gaussian process and proceeding using MCMC.
- A Hamiltonian-Monte-Carlo move is used to update the Gaussian process hyperparamters, in which both unlabelled and labelled data is used to make inference.
- This involves using Hamilton's physical equations of energy and momenta to efficiently explore the target probability distribution

$$\begin{aligned} \frac{d\mathbf{p}}{dt} &= -\nabla_{\mathbf{x}} H(\mathbf{x},\mathbf{p}) \\ \frac{d\mathbf{x}}{dt} &= \nabla_{\mathbf{p}} H(\mathbf{x},\mathbf{p}). \end{aligned} \quad (2)$$

- This proposed semi-supervised approach is compared with both empirical Bayes approaches (learning the hyperparamters using L-BFGS) and using Bayesian approaches which ignore the unlabelled data when making hyperparameter updates.
- Computation of both the likelihood and gradients in our model is computational intractable due to the large number of proteins present.
- Naïve inversion of the associated covariance matrix leads to computational scaling of $O((ND)^3)$, where typically $N \approx 10,000$ and $D \approx 60$.
- We can employ a tensor decomposition of our covariance, which allows fast extended Trench and Durbin algorithms for matrix inversion to be employed.

$$K = \sigma^2 I_{nD} + J_n \otimes A, \quad (3)$$

where J_n is a matrix of ones and A is a Toeplitz matrix.
- The computational cost of likelihood and gradient computations becomes $O(D^2)$, when these techniques are employed representing significant savings.

Results

- Our first example is a *Drosophila Melanogaster* (fly) embryos dataset.
- The posterior estimates of the noise parameters using both the labelled and unlabelled data is shifted right towards 0.
- This indicates that the noise parameters is smaller when solely using the labelled data. This is likely a manifestation of experimental bias, since it is reasonable to believe that proteins with known prior locations are those which have less variable localisations

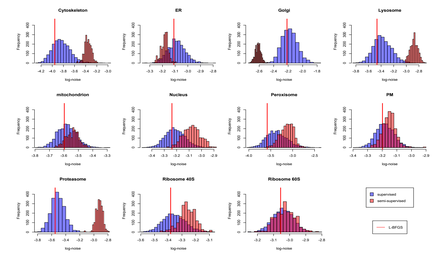

Figure 2:Posterior distributions for the log noise parameter σ^2.

- Furthermore, we notice enjoyable shrinkage in the posterior distribution of the noise parameter in the semi-supervised setting. The reduction in variance reduces our uncertainty about the underlying true value of σ_k^2 for $k = 1, ..., K$.
- Figure 3 demonstrates the results of applying our method. Each protein in this PCA plot is scaled according to mean of the Monte-Carlo samples from the posterior localisation probability.

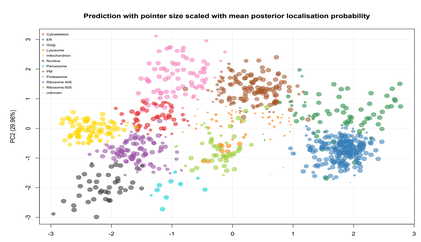

Figure 3:A pca plot for the *Drosophila* data where points, representing proteins, are coloured by the component of greatest probability. The pointer for each protein is scaled with membership probability.

Results

- Figure 4 highlights 3 proteins of interest and Figure 5 is a visualization of the probability that these proteins belong to specific classes.
- Figure 5 shows the cases of certain localisation, uncertain localisation between two classes, and no evidence of localisation to any sub-cellular niche.

Conclusion

- The proposed Bayesian framework performs consistently with previous methods whilst providing probabilistic information about protein sub-cellular localisations.
- This lays the foundation for more complex analysis including full estimation of the posterior assignment probabilities by Gibbs sampling and variational Bayes approximations.
- Further investigation is needed to fully exploit the potential of Bayesian models on spatial proteomics data.

Software

Code to perform the analysis on different datasets and to reproduce the analysis here is provided within the following Bioconductor packages

- MSnbase, pRoloc, pRolocdata

References

1. Christoforou, A et al.*A draft map of the mouse pluripotent stem cell spatial proteome* Nat. Commun. (2016)
2. Breckels, L et al. *Learning from Heterogeneous Data Sources: An Application in Spatial Proteomics* PLoS. Comp. Bio (2016)

Acknowledgements

Oliver Crook is a PhD student on the Wellcome Trust Mathematical Genomics and Medicine programme and acknowledges generous funding from the School of Clinical Medicine and the support of the Wellcome Trust. Laurent Gatto is head of the Computational Proteomics Unit and is funded by the BBSRC and the Wellcome Trust.